MODERN STATISTICS

IN PRACTICE

STEYN · SMIT · DU TOIT · STRASHEIM

Digitized by the Internet Archive
in 2025

https://archive.org/details/modernstatistics0000agws

MODERN STATISTICS
IN PRACTICE

A.G.W. STEYN
DEPARTMENT OF NATIONAL EDUCATION

C.F. SMIT
UNIVERSITY OF PRETORIA

S.H.C. DU TOIT
UNIVERSITY OF PRETORIA

C. STRASHEIM
UNIVERSITY OF SOUTH AFRICA

J L van Schaik
ACADEMIC

Published by JL van Schaik Publishers
1064 Arcadia Street, Hatfield, Pretoria
All rights reserved
Copyright © 1994 Authors

No part of this book may be reproduced or transmitted in any form or by any electronic or mechanical means, including photocopying and recording, or by any information storage and retrieval system, without written permission from the publisher.

First edition 1994

ISBN 0 627 01989 7

Cover design and illustrations by Barrett Joubert
Typesetting in 10½/12 pt Times by A1 Graphics, Pretoria
Printed and bound by Cape & Transvaal Book Printers, Parow, Cape

Hierdie boek is ook in Afrikaans beskikbaar onder die titel *Moderne statistiek vir die praktyk*.

BK4926

Contents

Chapter 1
Statistics: a means to an end

1 The role of statistics in society 1

2 The research process 3
- 2.1 Planning 3
- 2.2 Data collection 3
- 2.3 Editing and coding 3
- 2.4 Analysis 5
- 2.5 Conclusions 5

3 Different kinds of data 6
- 3.1 Quantitative data 7
- 3.2 Qualitative data 7

4 Data sources 8

5 The role of the computer in statistical analysis 9

6 Mathematical concepts 9
- 6.1 The sigma notation 9
- 6.2 The factorial notation 11
- 6.3 Natural logarithms 11

Exercises 12

Chapter 2
Sampling

1 Introduction 16

2 Problems with census taking 17
- 2.1 A census is expensive 17
- 2.2 A census takes a long time 17
- 2.3 The population is large 18
- 2.4 Sections of the population are inaccessible 18
- 2.5 Units are destroyed in the study 19
- 2.6 Inaccuracy of a census 19

3 Sampling error, sample bias and observation errors in sampling 19

4 Sampling methods 21
- 4.1 Probability procedures 21
  - 4.1.1 Simple random sampling 21
  - 4.1.2 Stratified random sampling 25
  - 4.1.3 Systematic sampling 30
  - 4.1.4 Cluster sampling 32
  - 4.1.5 Multistage sampling 35
- 4.2 Non-probability procedures 38
  - 4.2.1 Convenience sampling 39
  - 4.2.2 Judgmental sampling 39
  - 4.2.3 Quota sampling 39

5 Single sampling plans, multiple sampling plans and sequential sampling 40

6 Designing questionnaires 42
- 6.1 The content and phrasing of questions 43
- 6.2 Different types of questions 43
- 6.3 The appearance and layout of the questionnaire 45
- 6.4 Type of interview 46
- 6.5 Organizing the fieldwork 46
- 6.6 Editing and coding the data 47

Exercises 47

Chapter 3
Tabulation and graphic representation

1 Introduction 55

2 Arrays and frequency distributions 55

3 Frequency distributions of continuous data 58

4 Graphic representation of continuous data 62
- 4.1 The histogram 63
- 4.2 The frequency polygon 64

5 The cumulative frequency distribution and polygon 65

6 Relative frequency distributions 66

7 The cumulative relative frequency distribution and polygon 69

8 Frequency distributions and graphic representations of discrete and categorical data 70
9 Two-way frequency distributions and graphic representations 72
10 Using the computer for graphic representation 77
11 Concluding remarks 83
Exercises 83

Chapter 4
Descriptive measures of location or central tendency

1 Introduction 96
1.1 Descriptive measures of location 97
1.2 Descriptive measures of spread 97
1.3 Descriptive measures of symmetry 97
1.4 Descriptive measures of kurtosis 98
2 Measures of location: ungrouped data 99
2.1 The arithmetic mean 99
2.2 The weighted arithmetic mean 102
2.3 The median 103
2.4 The mode 104
2.5 The geometric mean 105
2.6 The quantiles 107
3 Measures of location: grouped data ... 109
3.1 The arithmetic mean 110
3.2 The quantiles 111
3.3 The mode 114
4 Concluding remarks 115
Exercises 116

Chapter 5
Measures of spread, symmetry and kurtosis

1 Introduction 123
2 Measures of spread: ungrouped data ... 124
2.1 The range 124
2.1.1 Properties of the range 125
2.2 The quartile deviation 127
2.2.1 Properties of the quartile deviation 128
2.3 Standard deviation 129
2.3.1 Properties of the standard deviation 133
3 Measures of spread: grouped data 134
3.1 The range and quartile deviation 134
3.2 The standard deviation 134
4 Interpretation of and uses for the standard deviation 137
4.1 Interpretation of the standard deviation 137
4.2 Relative measures of spread 139
4.3 Standardization 141
5 Moments 143
6 Measures of symmetry and kurtosis ... 146
6.1 Symmetry 146
6.2 Kurtosis 147
7 Exploratory data analysis 150
7.1 The stem-and-leaf display 150
7.2 Box-and-whisker plot 155
Exercises 156

Chapter 6
Fitting curves

1 Introduction 164
2 Why fit a curve? 168
2.1 Empirical determination of mathematical laws 169
2.2 Extrapolation 169
2.3 Interpolation 169
3 The linear relation 170
4 The coefficients of determination and correlation 174
4.1 The correlation coefficient r 175
5 Graphic representations of residuals and causal relations 180
5.1 Outliers 182
5.2 Influential observations 182
5.3 Causal relations 182
6 Curves that can be transformed to the linear form 183

6.1 The exponential curve (single-log transformation) 183
6.2 The power curve (double-log transformation) 186
6.3 The hyperbolic curve (inverse transformation) 188
7 The parabola 192
7.1 Adjusted coefficient of determination (R_a^2) 193
8 Growth curves 196
8.1 Modified exponential curve 196
8.2 Logistic curve 197
8.3 Gompertz curve 197
8.4 Determining parameters by means of the three-point method 200
9 The trivariate linear model 202
Exercises 207

Chapter 7
Time series

1 Introduction 219
1.1 The long-term movement (L) ... 220
1.2 Seasonal fluctuation (S) 221
1.3 The cyclical movement or business cycle (C) 221
1.4 Irregular variation (I) 222
1.5 The multiplicative model 222
2 Long-term movement 223
2.1 The method of least squares 223
2.2 The moving average method 227
3 Seasonal fluctuation 230
3.1 Calculation of specific and typical seasonal indices 231
4 Cyclic movement 236
5 Concluding remarks 239
Exercises 240

Chapter 8
Indices

1 Introduction 250
2 Different types of indices 252
2.1 Unweighted composite price indices 252
2.2 Weighted composite price indices 254
2.3 The weighted arithmetic mean of relative prices 256
3 Problems in determining and interpreting indices 260
3.1 The base period 260
3.2 Selection of items in composite indices 261
3.3 Shifting the base period 261
4 Important indices 262
4.1 The consumer price index of South Africa (CPI) 262
4.1.1 Base period 263
4.1.2 Selecting items and determining weights 263
4.1.3 Listing of prices 264
4.1.4 Calculation of index 264
4.2 The Johannesburg Stock Exchange indices 267
4.3 The Dow Jones industrial shares index 267
4.4 The intelligence quotient (IQ) ... 267
4.5 Other indices 268
5 Index chains 268
6 Deflation 269
7 Mathematical properties of indices 271
Exercises 273

Chapter 9
Introduction to probability theory

1 Introduction 280
2 Basic concepts 281
3 Rules of probability 288
3.1 The boundaries of probability ... 288
3.2 The addition rule 288
3.3 The complement rule 289
3.4 Conditional probability 291
3.5 The product rule 293
3.6 Statistically independent events 295
3.7 The law of total probability 297
3.8 Bayes' rule 299

4 Counting techniques 302
4.1 The basic counting rule 302
4.2 Permutations 304
4.3 Combinations 306
4.4 Methods of approximation 308
Exercises 310

Chapter 10
Probability models

1 Introduction 317
2 Basic concepts 317
2.1 Stochastic variables and associated distributions 317
2.2 Cumulative distributions 324
2.3 The mean and standard deviation of a stochastic variable 329
3 The binomial distribution 332
3.1 Tables of the cumulative binomial distribution 336
4 The Poisson distribution 336
4.1 The Poisson approximation of the binomial distribution 338
4.2 Tables of the cumulative Poisson distribution 339
5 The hypergeometric distribution 339
6 The exponential distribution 342
7 The normal distribution 344
7.1 Tables of the standard normal distribution 346
8 Sampling distributions and the central limit theorem 353
8.1 Statistics and sampling distributions 355
8.2 The central limit theorem 356
8.3 The normal approximation of the binomial distribution 358
9 The χ^2 distribution, the t distribution and the F distribution 362
9.1 The χ^2 distribution 362
9.1.1 Tables of the χ^2 distribution 362
9.2 The t distribution 365
9.2.1 Tables of the t distribution 366
9.3 The F distribution 367
9.3.1 Tables of the F distribution 368
10 Summary of sampling distributions 370
10.1 The one-sample problem 370
10.1.1 Pivotal quantity for μ 370
10.1.2 Pivotal quantity for σ^2 371
10.1.3 Pivotal quantity for π 371
10.2 The two-sample problem 371
10.2.1 Pivotal quantities for $\mu_1 - \mu_2$ 371
10.2.2 Pivotal quantity for $\frac{\sigma_1^2}{\sigma_2^2}$ 372
10.2.3 Pivotal quantity for $\pi_1 - \pi_2$ 372
Exercises 372

Chapter 11
Estimation

1 Introduction 381
2 Point estimation 382
2.1 The population mean μ 382
2.2 The population variance σ^2 383
2.3 The population proportion π 383
2.4 Characteristics of point estimators 384
2.4.1 Unbiasedness 384
2.4.2 Efficiency 384
2.4.3 Consistency 385
3 Interval estimation 387
3.1 Confidence interval for μ 387
3.2 Confidence interval for σ^2 389
3.3 Confidence interval for π 390
4 Determining sample size 393
4.1 Determining sample size when estimating μ 394
4.2 Determining sample size when estimating π 396
5 Closing remarks 398
5.1 Two-sample case 398
5.2 Finite correction factor 398
5.3 Table of point estimators and confidence intervals 398
Exercises 400

Chapter 12
Hypothesis testing

1 Introduction 405
2 Statistical hypotheses 407

3 Errors and risks 409
4 Test statistics and decision rules 414
4.1 Steps in hypothesis testing 416
4.2 p-values in hypothesis testing .. 419
5 The one-sample case 420
5.1 The population mean μ 420
5.1.1 Large-sample test for H_0: $\mu = \mu_0$ 421
5.1.2 Small-sample test for H_0: $\mu = \mu_0$ 422
5.2 The population proportion π 424
5.3 The population variance σ^2 426
6 The two-sample case 428
6.1 Difference between means (large samples) 428
6.2 Difference between means for small samples, given $\sigma_1^2 = \sigma_2^2$ 431
6.3 Difference between means of small samples, given $\sigma_1^2 \neq \sigma_2^2$ 432
6.4 Difference between means (paired observations) 434
6.5 Equality of proportions 440
6.6 Difference in variances 442
7 The power of a test 444
8 Concluding remarks 451
Guide to hypothesis testing 452
Exercises 453

Chapter 13
Linear regression and correlation

1 Introduction 472
2 Sample variance around the regression line 475
3 Inference regarding the intercept α 477
4 Inference regarding the slope β 482
5 Confidence band for $\alpha + \beta x$ 486
6 Prediction 489
7 Inference regarding the correlation coefficient ρ 491
7.1 Testing H_0: $\rho = 0$ 492
7.1.1 Tables for the r(n) distribution .. 492
7.2 Confidence interval for ρ (the z transformation) 495
7.3 Testing H_0: $\rho_1 = \rho_2$ 497
8 The trivariate linear model 498
9 Concluding remarks 500
9.1 The general linear model 500
9.2 Calculation aids 500
Exercises 500

Chapter 14
Analysis of variance

1 Introduction 508
2 The completely randomized design ... 510
2.1 Testing for one or more differences between a number of treatments 510
2.1.1 Notation 510
2.1.2 Test procedure 512
2.1.3 Calculation formulas 513
2.2 Scheffé's method of pairwise comparisons 517
3 The randomized block design 520
3.1 Notation 521
3.2 Calculation formulas 522
4 Factorial designs 526
4.1 Notation 528
5 Concluding remarks 533
5.1 Models for analysis of variance . 533
5.2 Stochastic effect models 533
5.3 Missing data 534
5.4 Validity of assumptions 534
5.5 Computer programs 534
Exercises 535

Chapter 15
Contingency tables

1 Introduction 547
2 The χ^2 goodness-of-fit test 549
3 The (2 × 2) table (fourfold table) 553
3.1 The (2 × 2) independence test 553
3.2 McNemar's test for two equally probable characteristics 556

4 The $(I \times J)$ table 559
4.1 The $(I \times J)$ independence test 559
4.2 Test for the homogeneity of independent samples 562

5 The log linear analysis of an $(I \times J)$ table 564
5.1 Decision between saturated and independence models 567
5.2 Estimation of the parameters for the appropriate model 568
5.2.1 Saturated model 568
5.2.2 Independence model 568

6 In conclusion 576
6.1 Multi-dimensional contingency table 576
6.2 Computer programs 576

Exercises 576

Chapter 16
Distribution-free methods

1 Introduction 583

2 The one-sample problem (testing the population median) 584
2.1 The sign test 584
2.1.1 Tables for the sign test 585
2.2 The test for symmetry 589
2.2.1 Tables for the signed rank test ... 591

3 The two-sample problem 594
3.1 Difference in medians (independent samples) 594
3.1.1 Tables for the rank sum test 595
3.2 Differences in medians (dependent samples) 597
3.3 Difference in spread 600

4 The multi-sample problem 603
4.1 The Kruskal-Wallis test (independent samples) 604
4.2 The Friedman test (associated samples) 607
4.2.1 The Friedman test for random block experiments 607
4.2.2 The Friedman test for the problem of h rankings 610

5 The problem of randomness 612

6 Rank correlation 615

7 Estimation 620
7.1 Point estimate of the median η 620
7.2 Interval estimation of the median η 620

8 Conclusion 622
8.1 Further distribution-free tests ... 622
8.2 Comparison of tests 622

Exercises 623

Appendices

Appendix A Data Set 1 629
Data Set 2 651
Data Set 3 667
Data Set 4 673
Data Set 5 675
Appendix B Statistical tables 677
Appendix C Weights of consumer goods 704
Appendix D Time series data 715

Index 728

Preface

This book was written with a twofold purpose. In the first instance, it was written for use in first-year courses in statistics. Students such as B. Com. first-years have all passed mathematics at matriculation level, but most of them have no interest in mathematically orientated subjects. The idea behind a service course of this nature is to acquaint students with important statistical techniques that are in daily practical use. We trust that this book will find application in such service courses. However, we do believe that the book can also be used as an introduction to more advanced courses in statistics, and that it can even serve as a basis for a course in Mathematical Statistics.

Research in virtually all disciplines has become increasingly dependent on the application of statistical techniques. Many researchers have little or no training in statistics and we believe that this book can serve as a source of reference for such individuals. One must, however, guard against a recipe-like application of statistical techniques by researchers who are not thoroughly conversant with the purpose, meaning and limitations of these techniques. Researchers who are not familiar with statistics are well advised to avail themselves of specialist statistical advice concerning the analysis of the data collected, before consulting textbooks on the application of specific statistical techniques.

Because the application of statistics is found primarily in daily practice, our approach throughout this book has been to explain statistical techniques by means of observed data sets. Where possible, we have used typically South African data sets and problems in the examples and applications. We hope by this means to impress on the reader the importance of the role played by statistics in this country. Readers will also be exposed to the rich variety of interesting facts that are continually being collected by various institutions in South Africa. The content of this book is, to a large extent, based on the knowledge and experience that we have gained through years of teaching, consultation and application of statistics in everyday life.

This book is divided into two sections. Chapters 1 to 8 contain a full overview of descriptive statistics. Application of descriptive methods is usually regarded as the first step in statistical inference. In Chapters 9 to 16 probability theory and distribution theory are focused on, as also the important techniques of statistical inference as required in the practical application of the statistical research process. Through these inferential techniques and based on descriptive measures calculated for the observed data set, conclusions are made regarding the population from which a sample has been drawn.

Despite the fact that the mathematical and theoretical levels of Chapters 9 to 16 are substantially higher than is the case in the first eight chapters, we have tried to emphasize the practical aspects of statistical techniques rather than the theory. However, important expansions on the subjects discussed, and also some of the more theoretical aspects, are briefly mentioned in notes. Readers who are interested can use the notes as a point of departure for a more detailed study of these subjects in the available literature.

The material in this book is definitely too extensive for an introductory semester course. For this reason, the book must be used selectively. Certain chapters can be omitted without any loss of continuity.

An Afrikaans version of the text has been available since 1981. We have recently identified a need for an English translation. We trust that this book will satisfy the need. There is a radical difference between earlier Afrikaans editions and this first English edition. Apart from the expansion of Chapters 2, 9 and 10, many of the examples have been replaced. The new examples are based on more recent and typically South African data sets. The five detailed data sets in Appendix A feature strongly in these examples. New exercises – far more than there were in previous editions – have also been included.

We gratefully aknowledge the valuable inputs of a number of people: Magdaleen du Toit and Dot van Wyk who did the editing in a very effective way, Linette Downes and Ernst Schlatter of J L van Schaik who respectively managed the whole project and produced both the Afrikaans and English versions and Mr George Louw who always gave most sympathetic support in difficult circumstances. A1 Graphics and Pace Setting and all other people behind the scenes who did the typing and took care of technical matters also deserve our grateful acknowledgement. Finally we want to mention our employers, namely the University of Pretoria and the University of South Africa, without whose support and encouragement we would not have been able to complete this project.

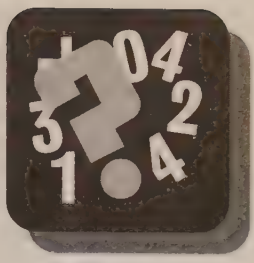
We trust that, even more than previous editions, this one will do justice to its title, namely that the content is modern and practically orientated.

A.G.W. Steyn
C.F. Smit
S.H.C. du Toit
C. Strasheim

July 1994

CHAPTER 1

Statistics: a means to an end



1 The role of statistics in society

If civilization is to progress, knowledge must be increased at all levels of society. Inventions, formulation of new theories, and refinement of existing processes are all heavily dependent on scientific research. Relevant data is collected, analysed and evaluated in terms of the scientific method of investigation. Statistics is an essential link in this process because it entails the development and application of techniques for interpreting numerical data. In short, one can say that:

Statistics makes sense of numbers.

The word statistics has various meanings. Uninformed people assume that statistics always refers to tables containing numbers on subjects such as population size, development of resources, and share prices. This interpretation of the word statistics is not necessarily wrong, but in this book we prefer to refer to such tables as **data sets**. We intend to use 'statistics' as a collective noun for the techniques that can be applied for analysis of data sets. In fact, over the past decade or two, the term 'statistics' has been used increasingly to describe the subject or discipline aimed at teaching these analysis techniques and at developing new techniques through basic research. One could add that, nowadays, all universities offer courses in statistics.

More than two thousand years ago the rulers of the Babylonian, Egyptian and Roman empires were already collecting 'statistics'. Such statistics usually contained information on the number of inhabitants in a country and on the nature and extent of its natural resources. Most civilized European countries have since the tenth century been drawing up registers of land ownership, births, deaths and marriages.

The few centuries preceding our own saw the emergence of precise sciences such as physics and astronomy which resulted in a greater emphasis on variation in measurements. Agriculturists and behavioural scientists began to study the interrelationship between various natural and genetic phenomena. However, it was only in the twentieth century that the study of collected information or data developed into an independent science with its own

character. Consequently this discipline, which is called statistics, is largely a product of the twentieth century; and aims to meet both contemporary and future research challenges. It is a dynamic discipline which devises methods for solving, *inter alia*, the following problems in diverse subject fields:

Economics: Determining which socio-economic variables (job creation, housing, health services, education) have the greatest effect on promoting economic growth.

Marketing: Determining what effect a comprehensive advertising campaign has on the sales figures of a new soft drink.

Epidemiology: Studying the incidence and distribution of AIDS and estimating the effect it will have on population figures over a twenty-year period.

Medical science: Determining the most important risk factors in the incidence of coronary thrombosis.

Demography: Estimating the average number of children per adult woman over a ten-year period based on certain trends that have been observed over the last five years during a population development programme.

Education and training: Estimating the number of qualified persons who will have to be trained in certain occupations in the next ten years in order to meet the manpower needs if the economy grows at an average rate of 1% per annum.

Agriculture: Identifying the variety of wheat best suited to a particular agricultural district in respect of factors such as yield per hectare and hardiness against drought and pests.

Insurance: Developing retirement annuity funds and calculating the contribution rates and retirement benefits of pension funds.

Engineering: Determining the twenty-year flood-mark when designing a bridge over a river.

Industry: Developing a suitable quality control system for electronic parts to ensure that the percentage of defects in manufactured parts does not exceed 0.1%.

Politics: Estimating the support enjoyed by a political party, a particular candidate or a particular policy in the run-up to an election.

Letters and words and the sequence in which they are placed are of no consequence to illiterates. Similarly the interpretation of figures or quantitative information, in other words statistics, requires a certain expertise. One needs to be statistically literate in order to read and interpret quantitative information or data sets with insight and understanding. However, in order to analyse data scientifically or to design experiments or surveys whereby data sets may be prepared, one has to be trained as a statistician. Because of its wide application

statistics creates many job opportunities for the qualified statistician. A graduate statistician can become a teacher or lecturer in his field. The public service, scientific councils, banks, insurance companies, financial institutions (such as mining) and industries, all require statisticians. Experience has also shown that statisticians who can critically and clinically analyse available data are eminently suited to the position of managers and decision makers at the highest levels of any job environment.

2 The research process

The accompanying schematic representation in Figure 1.1 is used to explain the role of statistics in research projects and other surveys or investigations. The striped circles in the diagram indicate the steps in the process where statistics plays a decisive role. The process can be divided into five steps or phases.

2.1 Planning

No important scientific research project can be conducted without thorough planning. The most important aspect of the planning phase is the formulation of the objectives of the investigation as precisely as possible. In order to achieve this specific goal information must be obtained from the population of items or persons being investigated. It is important to obtain only relevant information. If the project is to involve people and a substantial amount of information is required of each individual, questionnaires can be used to good effect. To save time and costs, the population items are not usually all investigated: only a sample is taken. The design of a suitable method of sampling is important at the planning stage and is discussed in Chapter 2.

2.2 Data collection

The actual collection of the data, in other words, the fieldwork, does not present many problems if the planning has been done carefully. This step can take as little as an hour or as long as several years, depending on the nature of the investigation and on the availability of funds and research assistance. In some laboratory situations the required data can be obtained within a few minutes. Against this a medical research project conducted among patients suffering from a rare disease may easily span ten years or more.

2.3 Editing and coding

It is important to edit all the data that has been collected, to eliminate obvious errors. Errors usually occur in research investigations where extensive data has been collected by numerous fieldworkers. Data that is of a dubious nature should preferably be eliminated, although this should not be applied dogmatically. When a substantial volume of data has been collected by means of questionnaires, the information is usually coded on the questionnaire and then captured on computer directly from the questionnaire. Editing is then done by computer.

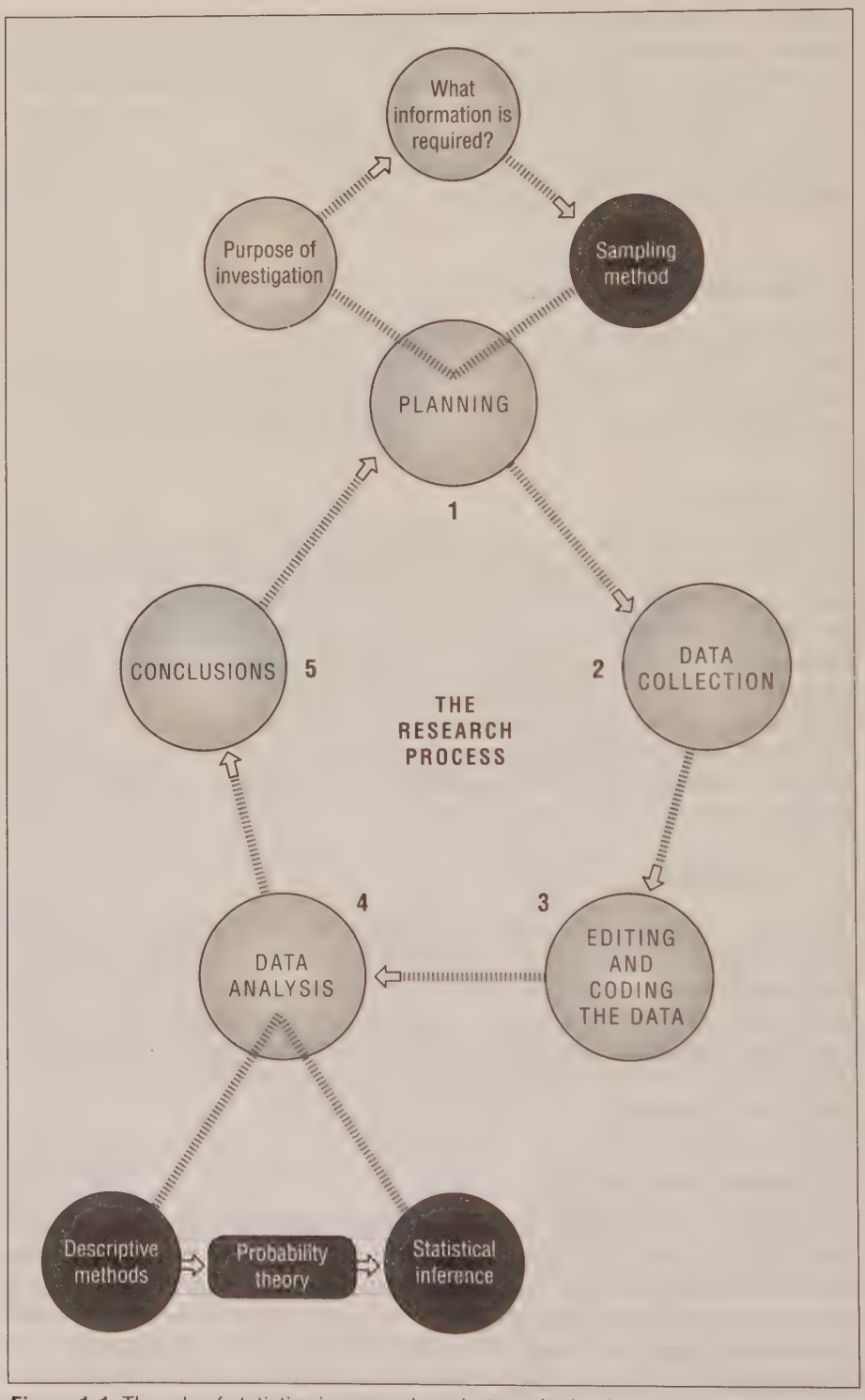


Figure 1.1 The role of statistics in research projects and other investigations

2.4 Analysis

At this stage of the investigation, statistical techniques are used to analyse the data that has been collected. The first step in this analysis is known as descriptive statistics.

> **Descriptive statistics** entails ordering and summarizing the data by means of tabulation and graphic representation and the calculation of descriptive measures. In this way the inherent trends and properties of the observed data emerge clearly.

The second step in data analysis is known as statistical inference or statistical generalization.

> **Statistical inference** draws conclusions about the population from which the sample was drawn by using the descriptive measures that have been calculated.

Information on a particular population will never be one hundred percent correct, unless the population is investigated fully.

Descriptive statistics and statistical inference comprise the two main subjects of this book and will be discussed in Chapters 3 to 8 and 11 to 16 respectively. The **theory of probability** forms the bridge between these two themes and its most important principles will be discussed in Chapters 9 and 10.

2.5 Conclusions

Once the statistics have been analysed, the most important conclusions can be made known in the form of a report. It is important to provide as many answers as possible to the questions raised in the objectives of the project. The statistician plays a very important role in interpreting the results and in drawing conclusions. Where it has not been possible to obtain answers for specific questions addressed in the survey, this fact should be frankly stated in the report. In fact this very statement often leads to new or follow-up investigations. The research process thus becomes a cycle, as indicated by the schematic representation. In order to illustrate the research process we shall briefly discuss two examples.

EXAMPLE 1.1

In order to gauge the effect of an advertising campaign on the sales of a new product, the following experiment is carried out. Initially such a campaign is used to publicize the product in Pretoria. This entails the use of local newspapers, regional transmitters of the SABC and local cinema theatres for advertising space and time. Simultaneously the new product is offered for sale in shops in Port Elizabeth without any promotion campaign whatsoever. A representative sample of ten shops is taken in Pretoria and the total sales of the product at each of these shops is determined for the first month following the advertising campaign. Ten

shops of comparable size are then selected in Port Elizabeth and the total sales during the first month of the availability of the same product is determined. By means of statistical techniques it is now possible to ascertain whether the sales of this product have been higher in the Pretoria shops than in the Port Elizabeth shops. If such a conclusion can be substantiated by sales figures, the positive effect of an advertising campaign will have been established. This type of conclusion can only be made by means of statistical inference. The correct selection of the ten shops from each of the two cities is of critical importance in the investigation and is a specialized task requiring the assistance of a statistician.

EXAMPLE 1.2

The municipality of a certain city envisages decentralization of municipal services. In order to identify the areas with the greatest need for new branch libraries, medical clinics, waste dumps and administrative centres where, for example, bus tickets could be purchased and accounts be paid, the authorities decide to conduct an opinion poll among city residents. A questionnaire is designed to assess the residents' preferences regarding these municipal services. To save costs, this questionnaire is sent to a representative sample of 10% of all home- and flat-dwellers by including it with the monthly municipal account. Despite the fact that only 30% of the inhabitants thus selected have returned their completed questionnaires after a period of three months and an additional 10% of these questionnaires have had to be discarded after editing, the city council still decides to proceed with the investigation. After capturing the collected data onto computer, the data is then analysed statistically. By using these analyses the city council can now decide which services justify further decentralization; and in which suburbs these services must be made available to the public. It is obvious that the planning step in this investigation was very important. The questionnaire had to be compiled without unnecessary prolixity but with sufficient attention to detail to obtain all the required information from residents. It was important to ensure that residents from every suburb, age group and income group were properly represented in the sample. All these aspects will be discussed in Chapter 2.

3 Different kinds of data

Schematically, the different kinds of data can be represented as follows:

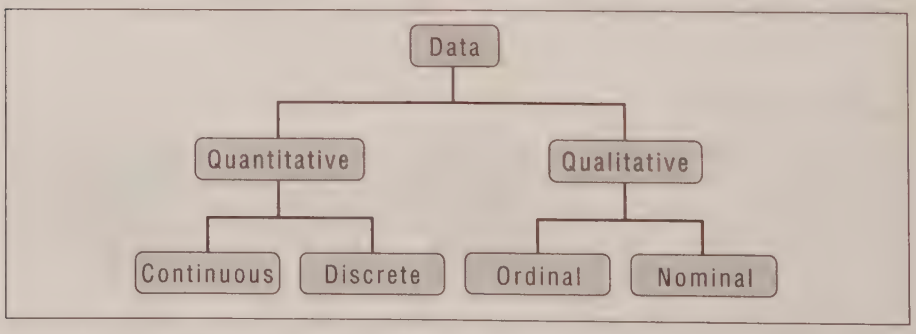


Figure 1.2 Schematic representation of the different kinds of data

3.1 Quantitative data

If any characteristic or property that is particular to a person or entity is measurable on a numerical scale, information on such a characteristic or property is called **quantitative data**. Thus, for example, information regarding people's age and income and information regarding motor cars' fuel consumption and year model is said to be quantitative. Quantitative data is usually classified as either discrete or continuous.

> If observations of a particular characteristic can assume only fixed isolated values, this data is known as **discrete data**.

Information on the number of bedrooms in houses purchased is an example of discrete data, since each house can have only 1, 2, 3, 4, 5 or 6 bedrooms. Other examples of discrete data are: the numbers that may possibly come up when a hexagonal dice is repeatedly thrown; the numbers to be found on the jerseys of members of a rugby team; and the number of people per family unit.

> If observations of a particular characteristic can assume all possible values within a particular logical interval, such data is called **continuous data**.

Data on the floor area of the buildings of houses purchased is an example of continuous data. Other examples are the ages, mass, height and salaries of men; and the fuel consumption of motor cars.

Obviously, most quantitative data is of a continuous nature. Although age is usually indicated in years and annual salary to the nearest thousand or hundred rand, this information should not be regarded as discrete data. Age can in fact be determined to the nearest minute and salary to the nearest cent, but rounding to appropriate units makes such data easier to handle.

3.2 Qualitative data

Many of the characteristics that are inherent in a person or entity cannot be reflected numerically. Data that is not numerical would include: information on the colour of a person's eyes or hair, information on a person's attitude towards his church or work; and information on the subjects that a student includes in his degree course. This information is known as **qualitative** or **categorical data**. However, qualitative data can often be represented in a quantitative manner. Thus an employer can quantify his employees' attitude towards their work by describing it as 'poor', 'fair', 'good' or 'excellent' and then assign rankings to these descriptions on a scale of, say 0 to 3, namely 0, 1, 2 and 3 respectively. In this example there is a definitive order in the categories and one would refer to **ordinal data**, in contrast with, for example, information on the colour of a person's eyes. Such information can be classified as **nominal data**. Note that the figures 0, 1, 2 and 3 have no physical meaning; they are used only

to rank attitude. This is definitely not an indication that an excellent attitude is three times more positive than a fair attitude. Conversely, it is also possible to divide quantitative data into categories, such as, for example, age, which can be represented as 'young', 'middle-aged' or 'old'.

It is essential to distinguish between the different types of data when choosing appropriate statistical techniques, because the technique to be applied is often determined by the kind of data that is to be analysed. A large data set often contains more than one type of data. An example is a questionnaire in which information on the respondent's age and marital status is requested in addition to his attitude or views on matters that are related to the main topic of the questionnaire.

4 Data sources

Collecting the data is an important component of any research project, but can often be time-consuming. If information is collected from scratch, the sources used are known as **primary data sources** and the data as **primary data**.

However, in many cases the required data is already partially or fully available and need neither be collected physically by fieldworkers nor be obtained through laboratory experiments. Such data is known as **secondary data**. To save costs and time, data sources that are readily available should be identified and used.

The most comprehensive data sources in South Africa are the publications and databases put out by the Central Statistical Service (CSS), which has its headquarters in Pretoria. Publications such as *SA Statistics* (biennial) and *SA Labour Statistics*, which can be purchased from the Government Printer and are available in libraries, provide comprehensive information on demography, prices, production, rainfall, et cetera. The CSS also publishes, monthly and quarterly, an extensive number of news releases containing the latest information on topics such as consumer prices, production prices, building costs, et cetera. Much of this information is provided in the form of indices. Comprehensive information reports on all facets of formal education are disseminated annually by the Department of National Education.

Annual reports of government departments, municipalities, companies and research councils are often useful sources of information for more specialized information. Publications of the World Bank, the United Nations Educational, Scientific and Cultural Organization (UNESCO), and the Organization for Economic Co-operation and Development (OECD) offer a wealth of economic, educational and other related information on many countries. Information from these sources is especially useful in international comparative studies. Newspapers, too, are data sources and information that is extracted from these sources can also be of great value provided it is used judiciously. Organizations and institutions should realize the importance of establishing data banks and libraries containing relevant information and should supplement and update these on an ongoing basis. In statistical analysis it is common practice to use both primary and secondary data sources.

5 The role of the computer in statistical analysis

Phenomenal developments in the field of electronics over the past few decades have greatly facilitated the practical use of virtually any statistical analysing technique, however complex. A pocket calculator will process complicated calculations within seconds, while computer programs for sophisticated statistical techniques can be purchased relatively inexpensively for use on both mainframe and personal computers. However, a word of warning is due here. Although individuals without any knowledge of statistics can quite easily apply a variety of statistical techniques by means of these programs, the results of the analyses cannot be interpreted without a thorough understanding of statistics. The emphasis in teaching statistics is therefore shifting further and further away from the aspect of calculations toward the aspects of understanding, interpretation and use. Some of the best-known programs for statistical analysis that are used internationally and locally among educational and research institutions are:

Biomedical Computer Programs, P-series (BMDP)
Statistical Analysis System (SAS)
Statistical Package for Social Sciences (SPSS)
Statgraphics
CSS Statistica
SYSTAT
LOTUS 123
QUATTRO PRO
EXCEL
Harvard Graphics

These programs are continuously being improved and expanded. They are accompanied by comprehensive user manuals. Although these manuals usually provide brief descriptions of the statistical techniques that are used in the programs, they are not statistical textbooks. These manuals can therefore only be properly used and appreciated by people who have a thorough understanding of the statistical analysis techniques applied in the programs.

6 Mathematical concepts

Although a basic knowledge of mathematics is taken for granted in the next chapters, it is important to discuss briefly certain mathematical concepts that may not generally be well known.

6.1 The sigma notation

Since the calculation of addition is so important in statistical techniques and will be used frequently in this book, a brief notation for this calculation, known as the **sigma notation**, will be discussed. In symbols the sum of n numbers can be expressed as follows:

$$x_1 + x_2 + \ldots + x_n = \sum_{i=1}^{n} x_i$$

Σ is the Greek capital letter sigma (the equivalent of the Latin letter S), and its meaning in the above context is 'sum of'. The index i assumes all integer values from 1 to n.

EXAMPLE 1.3

(a) $x_1 + x_2 + x_3 = \sum_{i=1}^{3} x_i$

(b) $x_3 + x_4 + x_5 + x_6 = \sum_{i=3}^{6} x_i$

(c) $y_5 + y_6 + y_7 + y_8 = \sum_{i=5}^{8} y_i$

The following two basic rules apply when this notation is used:

Rule 1

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

Rule 2

$$\sum_{i=1}^{n} cx_i = c \sum_{i=1}^{n} x_i$$

By letting $x_1 = x_2 = \ldots = x_n = 1$ and applying Rule 2 it follows that

$$\sum_{i=1}^{n} c1 = cn$$

EXAMPLE 1.4

(a) $\sum_{i=1}^{n} (ax_i + by_i) = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} y_i$

(b) $\sum_{j=1}^{k} (y_j - 3)^2 = \sum_{j=1}^{k} (y_j^2 - 6y_j + 9)$

$$= \sum_{j=1}^{k} y_j^2 - 6 \sum_{j=1}^{k} y_j + 9k$$

Note

With the repeated use of the sigma notation, and if no possibility of confusion on the summation limits exists, a simplified notation can be followed, namely:

$$\sum_{i=1}^{n} x_i = \sum x$$

In terms of this simplification, rules 1 and 2 can be written respectively as:

$$\sum (x + y) = \sum x + \sum y$$
$$\sum cx = c \sum x$$

6.2 The factorial notation

This notation offers a concise method of expressing the product of a particular collection of integers:

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1$$

EXAMPLE 1.5

(*a*) $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$

(*b*) $$\frac{n!}{(n-k)!} = \frac{n \times (n-1) \times \ldots \times 2 \times 1}{(n-k) \times (n-k-1) \times \ldots \times 2 \times 1}$$
$$= n \times (n-1) \times \ldots \times (n-k+1)$$

In order to ensure that this notation applies consistently, it is accepted by definition that:

$$0! = 1$$

Consequently in Example 1.5(b) with $n = k$, it follows that $(n - n)! = 0! = 1$.

6.3 Natural logarithms

Despite the fact that, before the advent of the computer, the use of logarithms to the base of 10 led to simplified calculation of products, quotients and exponents, the natural logarithm (or logarithm to the base of e) is mathematically more important.

If $y = e^x$

where $$e = \sum_{i=0}^{\infty} \frac{1}{i!}$$
$$= 2.7182818285 \ldots$$

then $\log_e(y) = x$

Usually $\log_e(y)$ is abbreviated as $\ell n y$ and is known as the **natural logarithm** of y.

The rules that apply with regard to logarithms to the base of 10 also apply here, namely:

$$\ell n(a.b) = \ell n a + \ell n b$$

$$\ell n\left(\frac{a}{b}\right) = \ell n a - \ell n b$$

$$\ell n a^b = b\ell n a$$

$$\ell n e = 1$$

Only logarithms to the base of e will be used in the remaining chapters.

All standard pocket calculators are equipped with the natural logarithmic function. The following can therefore be calculated easily:

$$\ell n 5 = 1.6094379$$
$$\ell n 10 = 2.3025851$$
$$\ell n 2.718281 = 0.9999997.$$

Exercises

1 Simplify the following (see Example 1.4):

(a) $\sum_{i=1}^{n}(4x_i - 2)$

(b) $\sum_{k=2}^{5}(w_k + 4)$

(c) $\sum_{i=1}^{m}(6x_i + 3y_i)$

(d) $\sum_{l=1}^{t}(p_l + 5)^2$

(e) $\sum_{j=1}^{n}(cz_j + dv_j^2)$

(f) $\sum_{i=1}^{m} 3a_i^3$

(g) $\sum(3z + 4y)^2$

(h) $\sum_{j=1}^{3} y_j - \left(\sum_{i=1}^{3} y_i\right)^2$

(i) $\sum 4x(x + 2)$

(j) $\sum(y + 2y^2)^2$

(k) $\sum_{k=1}^{3}(w_k - 1)(2w_k + 3)$

(l) $\sum(4d^2 - q)$

(m) $\sum_{s=1}^{4}(x_s + 4)$

(n) $\sum_{i=1}^{5}(3x_i - i)^2$

2 Assume that $x_1 = 3,\ x_2 = -7,\ x_3 = 2,\ x_4 = 10,\ x_5 = 6$
$y_1 = 11,\ y_2 = 10,\ y_3 = 12,\ y_4 = 15,\ y_5 = 14$
Compute the following:

(a) $\sum_{i=1}^{4} x_i$ (b) $\sum_{j=2}^{4} y_j$

(c) $\sum_{i=1}^{5} (3x_i - 2y_i)$ (d) $\sum_{k=1}^{4} 2y_k^2$

(e) $\sum_{l=1}^{4} (4y_l^2 + x_l^2)$ (f) $\sum_{j=1}^{3} (bx_j + ay_j)$

(g) $\sum_{i=1}^{4} x_i y_i^2$ (h) $\sum_{j=0}^{2} x_j y_j$

(i) $\sum_{i=1}^{5} \sqrt{y_i}\ x_4$ (j) $\sum_{i=3}^{5} x_{i-2} y_{i-1}$

3 Are the following expressions correct or incorrect? (Give reasons if you think they are incorrect.)

(a) $\sum \left(\frac{1}{2} x\right) = \frac{1}{2} \sum x$

(b) $\sum \frac{x}{y} = \frac{\sum x}{\sum y}$

(c) $\sum x - y = \sum (x - y)$

(d) $\sum x^2 = (\sum x)^2$

(e) $\sum (x^2 - y^2) = \sum x^2 - \sum y^2$

(f) $\sum (cx + y) = c \sum (x + y)$

(g) $\sum x.y = \sum x \sum y$

(h) $(\sum x + \sum y)^2 = \sum x^2 + 2 \sum x \sum y + \sum y^2$

4 If $\sum_{i=1}^{k} i = \frac{1}{2} k(k + 1)$, express the following in terms of the sigma notation and use these results to do the following calculations:

(a) $1 + 2 + 3 + 4 + 5$

(b) $5 + 6 + 7 + \ldots + 19$

(c) $n + (n + 1) + \ldots + (n + k)$

5 Calculate:

(a) $\frac{20!}{16!}$ (b) $\frac{7! \times 5!}{35 \times 24 \times 6}$

(c) $\sum_{k=1}^{4} 5k!$ (d) $\sum_{j=0}^{5} (j + j! + 2)$

(e) $\sum_{l=0}^{3}(l+4)!$ (f) $\dfrac{\left(\sum_{j=1}^{4} p\right)!}{\sum_{j=1}^{4} p!}$

6 Calculate the value of e accurately to 15 decimals.

7 Determine the logarithm of y:

(a) $y = 4e$ (b) $y = e^3$

(c) $y = e^{x^2}$ (d) $y = \dfrac{e}{5}$

(e) $y = 2e^{x+v}$ (f) $y = \dfrac{10}{x^{\ln 10}}$

8 The heights (cm) of 10 adult mixed twins are the following:

No (i)	Men (x_i)	Women (y_i)
1	174	159
2	162	164
3	155	155
4	157	158
5	180	178
6	185	173
7	168	162
8	163	169
9	176	172
10	178	170

(a) Express the following in sigma notation without calculating the expressions.

(i) The sum of the heights of the 10 men.

(ii) The sum of the heights of the 10 women.

(iii) The difference between the sum of the heights of the 10 men and the sum of the heights of the 10 women.

(iv) The sum of the heights of the last 4 men.

(v) The difference between the sum of the heights of the first 5 men and the sum of the heights of the first 5 women.

(vi) The sum of the differences in height (man's height minus woman's height) of the 10 pairs of twins.

(b) Use the statistical function on your calculator to its maximum potential to calculate

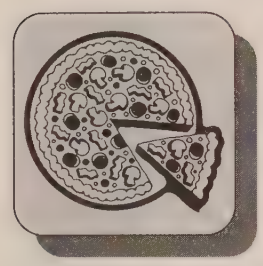
(i) $\sum x$ (ii) $\bar{x} = \dfrac{\sum x}{10}$ (approximate to 2 decimals)

(iii) $\sum x^2$ (iv) $\sum x^2 - 10x^2$

(v) $\sum x^2 - \frac{(\sum x)^2}{10}$ (vi) $\sum (x - \bar{x})^2$

(Note the different answers for (iv), (v) and (vi), which are mathematically speaking, identical. How can these differences be accounted for? Refer to Chapters 4 and 5 as well.)

(c) Express in sigma notation and calculate:
- (i) The sum of the squares of the heights of the 10 men.
- (ii) The sum of the heights of the 10 men, squared.
- (iii) The sum of the squares of the heights of the 10 women.
- (iv) The sum of the heights of the 10 women, squared.
- (v) The sum of the products of the heigths of the 10 men and those of the 10 women.
- (vi) The product of the [sum of the heights of the 10 men] and [the sum of the heights of the 10 women].



CHAPTER 2

Sampling

1 Introduction

Sampling is common practice and we are all continually being confronted with some or other aspect of sampling. 'Public opinion' on topical matters is often made public by the media after a quick opinion poll. From childhood we are actually all samplers. The first taste of a morsel of food often determines whether we intend eating the rest, and a quick glance through a book, whether we intend to read it.

Samples like these are not always representative of the aggregate from which they are taken. Ignorance and misconceptions regarding sampling abound, and, since it forms the basis of many statistical analyses of primary data sources, a relatively full discussion of the topic is appropriate. A suitable sampling method can only be selected once the researcher has acquired a complete description of the population from which the sample is to be drawn.

The **population** (also called the target population) is defined as the total group of persons or universal collection of items (elements) to which the study relates.

A **sample** is any subset of the population. If a total population is studied, we refer to a full study or a **census**.

The researcher uses the information obtained from the sample as a basis for generalizations on the entire population. The accuracy of these generalizations will depend on the accuracy with which the sample reflects the characteristics of the population. Thus the percentage of support enjoyed by a political candidate in a constituency can be estimated relatively accurately on the basis of a fairly limited but well-planned sample of the voters. Quality control of manufactured products is totally dependent on the correct use of sampling techniques. The success of any business is usually determined by the correct assessment of client preferences, for which sampling is essential.

The composition and nature of the population, and any additional information on the population, are among the most important factors that influence the choice of a specific sampling method. This chapter provides an overview of the most important methods. All these methods are simple to apply in practice and some of them could well be used more frequently.

2 Problems with census taking

A substantial section of this book deals with statistical inference, or statistical generalization. If information is obtained by means of a census, inference is not applicable, since a census provides comprehensive information with respect to the population and all its characteristics. However, there are various reasons why it is not possible to take a census to establish a primary data source and why sampling is then essential. We shall now discuss the most important of these reasons:

2.1 A census is expensive

On 7 March 1991 a population census was taken in South Africa. The purpose was to establish the number of inhabitants in the country, but information was also obtained on ethnicity, language, religion, income, occupation, et cetera. This census cost in the region of R65 million, partly because millions of questionnaires had to be printed and thousands of temporary personnel had to be employed to travel millions of kilometres to collect the information. Because of the enormous cost involved a population census is taken only once every five years. Despite the careful preplanning and high costs that are incurred, a population census yields an undercount of the population and the information has to be corrected on the basis of additional information. Not all censuses have the same scope as a population census, but usually they are expensive and they do cause considerable administrative problems.

2.2 A census takes a long time

It takes years to plan a population census. Distribution of questionnaires starts a month before census day, and it can take months to collect them all again. The census is rounded off by editing, processing and publishing of relevant information, but these steps are all time consuming and can take longer than a year. If information is needed quickly, a census is out of the question and sampling has to be used. A company can show a large profit if it determines the popularity of a new product at an early stage with the aid of a marketing survey, and then plans its marketing accordingly. A candidate in an election can strengthen his position if the election strategy is planned with due regard for regular and well-designed opinion polls among the voters. These surveys are worthless if they are not done speedily and accurately.

2.3 The population is large

A population always comprises a finite number of elements, even if the number is extensive. If a census cannot be taken within a reasonable time, sampling has to be used. It is, for example, practically impossible to take a census of all earthlings within a reasonable time. The information that is available on this population is inadequate and inaccurate and is based on internal censuses that have been taken in all the countries of the world. Some of these are not very reliable. The availability of time and money plays a decisive role in the choice between a census and a sample. These three factors, namely size of population, time and cost should therefore be considered jointly when a survey technique for any research project is being planned. In effect this usually means that a census cannot be regarded as cost effective.

2.4 Sections of the population are inaccessible

Certain parts of the world are so isolated that they can be reached only at great peril and cost. With nature surveys, for example, it is difficult to study animals and plants on mountains that are higher than 7 000 metres above sea level. Similarly, it is virtually impossible to study marine life at depths greater than 7 000 metres. Consequently, thousands of species living in the sea are still completely unknown. Access to very important persons, or persons in hospitals and prisons is often forbidden. By sheer necessity sections of the population therefore have to be excluded from a study. Even if the remainder of the population is then included, this section, however big, represents a subset and, technically speaking, only a sample of the population therefore has been drawn. Alternatively, sections of the original population that it is impossible to include can be excluded. This means that the target population diminishes in size. If the reduced population is still too large for a census, a sample from this population would be the obvious method. In Third World societies people often live in informal housing or squatter camps. Due to the lack of organization in such residential areas censuses there are not only difficult, but also inaccurate. Aerial photographs are increasingly being used in population censuses in these areas. Certain areas are selected for this purpose and sampled comprehensively. The information thus obtained is then combined with the aerial photographs and used to estimate census data for these areas.

Another problem that often surfaces is the poor response from questionnaires sent out to a population of persons or families. Because only a section of the population completes and returns the questionnaires, the proposed census then becomes a self-selected sample. Many experts believe that a response of 30% can be regarded as normal in such studies. Psychologists also maintain that respondents and non-respondents usually differ with regard to character traits. An analysis of the data collected will therefore not reflect a true picture of the total population. To bridge this problem, follow-up studies are often done and non-respondents are then contacted either telephonically or personally. Naturally this is an expensive and time-consuming process.

2.5 Units are destroyed in the study

One of the most important fields for the application of sampling and the accompanying statistical inference is quality control. To ensure maintenance of high quality in items produced the end product must be examined constantly. To obtain information on the quality of the product, a certain number of selected items are examined, and the decision as to whether or not the production process is still under control, that is whether the items produced still comply with the specifications, is then made on the basis of this sampling information. The items that are tested are often destroyed during testing. For example, when the life-span of a light bulb, fuse or electronic component has to be determined, a sample of these items is subjected to so-called accelerated experiments that provide information on the basis of which their actual life-span can be determined. If there is a possibility that expensive components used during testing can be destroyed the sample taken must obviously be a very small, representative sample.

2.6 Inaccuracy of a census

From the above discussions it is clear that due to practical, administrative and cost considerations a census can be problematic. The fieldworkers involved in the census also have to be taught and motivated beforehand, and during the census they have to be continually checked. If a census is not planned, carried out and administered with great expertise and enthusiasm, it can easily become a fiasco. For these reasons, a well-planned sample that is easy to administer often provides more reliable information than a full population census. Sampling is quicker than census taking and consequently fieldworkers maintain the required interest in sampling. Finally, the smaller volume of data involves less editing and processing, and the possibility of gross errors is therefore also smaller.

3 Sample size, sample bias and observation errors in sampling

In order to study the mean selling price of residences in the urban areas of South Africa, a sample of n houses is drawn according to a certain procedure. If other samples of the same size were to be drawn according to the same procedure, the average selling prices calculated for the different samples would differ. The differences between these sample averages as well as the differences between the actual population mean and the sample estimates of this mean are reflected in the sampling error.

The **sampling error** is caused by the inherent inability of a sample to provide totally accurate information on the population. This error is usually expressed in terms of the standard error of estimation (see Chapter 11).

Each sampling procedure has a corresponding sampling error and the sampling error also depends on the value of n, the sample size. To a large extent, it is impossible to control the sampling error, since it is determined largely by factors that are beyond the control of the sampler, but it can be limited by increasing the sample size. In the above example a relatively large number of luxury houses could accidentally be included in a small sample, with the result that an unrealistically high estimate of the mean selling price is made.

Sample bias is a phenomenon that occurs if sample elements are obtained in such a way that, for identifiable reasons, certain sections of the population are either not represented at all or represented only to a lesser degree.

If there is sample bias, the characteristics of the population will not be clearly reflected in the sample. For instance, the characteristics of non-respondents will not be reflected in a sample based on a questionnaire. A prime example can be found in the market research on views held by consumers of frozen vegetables concerning both fresh and frozen vegetables (see Appendix A, Data Set 2). The survey was conducted in supermarkets throughout the country and if the information was collected on weekdays when mainly non-working women go shopping, sample bias can be expected in the data. To a large degree, working women and men would have been excluded from this survey. Since working persons are often grateful for the convenience of frozen vegetables, the conclusions drawn from the observed sample with respect to the knowledge about and use of fresh and frozen vegetables could well be biased. In contrast to sampling errors, however, sample bias can be eliminated or minimized by thorough planning of the sampling procedure.

Finally, it is important to be aware that a third type of error can also emerge in the collection of sample data:

Sampling observation errors (also called non-sampling errors) are not the result of a certain sampling procedure or sample size, but are the errors made by the sampler or respondent during the collection of the sample data.

Erroneous or inaccurate measuring instruments, unreliability of either the interviewer or the respondent, ambiguous questions and erroneous or unclear written responses can all lead to this type of error. Experience has shown that answers on age, body mass and income, should usually be taken with a pinch of salt.

However, such information can often be obtained indirectly. It is an unfortunate fact that, due to sampling observation errors, investigations in which special care has been exercised to minimize sampling error and bias, often still yield worthless or inaccurate information. Well-planned questionnaires, well-trained fieldworkers and the use of reliable equipment in research surveys ensure that errors from this source are minimized.

4 Sampling methods

Sampling methods can be divided into two groups, namely probability procedures and non-probability procedures. The latter are far more subjective and should be applied circumspectly in important research projects.

4.1 Probability procedures

If each element in a population has a known probability or chance of being included in the sample, the sample is called a **probability sample**.

Simple random sampling, stratified random sampling, systematic sampling and cluster sampling are all probability procedures. Simple random sampling is the most important method, since its principles are also utilized in each of the other three methods.

4.1.1 Simple random sampling

The following over-simplified situation can be used to explain the application of this technique: In a survey on the income of employees in the building industry, a sample of the employees of each firm has to be drawn. One small construction firm employs five people and a simple random sample of size 2 is drawn from the employees of the firm. In this case the population therefore consists of five elements. If the employees are designated as A, B, C, D and E, the following ten samples are possible:

$\{A, B\}$; $\{A, C\}$; $\{A, D\}$; $\{A, E\}$, $\{B, C\}$;
$\{B, D\}$; $\{B, E\}$; $\{C, D\}$; $\{C, E\}$; $\{D, E\}$.

If all have an equal chance of being selected, then each of the ten samples would be a simple random sample. A simple random sample can therefore be drawn by writing a code for each of the ten samples $\{A, B\}$, $\{A, C\}$, ..., $\{D, E\}$ on ten identical cards, placing them in a container, shuffling them thoroughly and then drawing one of the cards. Obviously, it is not possible to apply this method in the case of large populations and samples.

The same result can be achieved by writing the letters A, B, C, D and E on five identical cards, placing the cards in a container, shuffling them thoroughly

and then drawing two consecutive cards. It is clear that each employee has an equal chance of being selected in the first draw. Once the first employee has been drawn, each of the four remaining employees has an equal chance of being selected as the second element in the sample. If the sampling is done in this way, we say that the selection of consecutive elements is mutually independent. The following definition of a simple random sample can therefore be given:

> A **simple random sample** is obtained if each element of the population that has not yet been included in the sample stands an equal chance of being selected in the next draw.

Due to man's inherent bias, no individual can draw a completely random sample unless some or other objective mechanism is used. However, no such mechanism can be used unless a complete list of all the population elements is available.

The method described above, where each population element is identified by a code on a card and the cards are drawn from a container after having been shuffled, is very cumbersome if the population is extensive. For this reason we use random digits when taking a simple random sample.

> **Random digits** are a random arrangement of the digits 0, 1, 2, ..., 9.

These digits can be generated in different random ways, for example by rolling a ten-sided dice, or by means of a mathematical algorithm on a computer. Ultimately, the digits 0, 1, 2, ..., 9 occur an equal number of times and there is no pattern in the sequence of digits.

Random digits are usually given in groups, each consisting of two or three digits. These groups are then arranged into rows and columns so that they can be easily read by the user. Consecutive digits (in the same row) can be grouped together as required to form, for example, 2 or 3 digit numbers. A table with 10 000 random digits is given as Table B.1 in Appendix B. Many pocket calculators these days have a built-in function for generating two- and three-digit random numbers, which makes the tables with random digits largely superfluous. For the sake of clarity in examples and uniform answers in class exercises, we will mainly use the random digits given in Table B.1 in this book.

EXAMPLE 2.1

Suppose we wish to draw a simple random sample of size 100 from the population of 843 selling prices of residences sold in Johannesburg from October to December 1991 (see Appendix A, Data Set 1). If the random digits in Table B.1 are used, we can proceed as follows:

(a) Select a 'random' starting point in Table B.1. Any point can be used, but with repeated use of the table care must be taken to choose different starting points. Avoid always starting at the top of a page or at the beginning of a row or column of digits. Suppose we decide to start on the third page of Table B.1, with row 16, column 29.

(b) Read systematically from that point in any direction. Reading from left to right in a horizontal line is a completely acceptable method. This method can, on occasion, be alternated by reading down vertically in adjacent columns, and by other methods.

(c) Divide the consecutive single digits into groups with the same number of digits as the population size and write down each of the numbers thus formed, that are smaller than or equal to the population size. Since $N = 843$ in the current example, the random digits are grouped into three-digit numbers and only those that are smaller than or equal to 843 are noted.

(d) The population elements with numbers that agree with the numbers obtained are included in the sample. The procedure is continued until the desired number of sample elements has been drawn, and any number that has already been used is ignored in any subsequent draw.

The 100 numbers that are obtained if the numbers are read horizontally from left to right from the starting point (third page, row 16, column 29), are the following:

92	365	411	403	662	759	316	776	612	533
174	275	687	21	161	372	281	812	159	40
594	180	359	782	266	711	753	86	480	231
658	118	20	805	691	447	557	656	700	15
833	728	716	320	780	496	333	141	609	132
826	134	173	378	113	186	409	284	89	261
714	80	620	154	808	827	30	683	834	720
450	103	448	727	273	830	758	108	263	766
416	626	628	692	214	523	306	553	434	396
315	493	596	509	460	840	805	763	546	788

The simple random sample of 100 selling prices of residences was taken from Data Set 1 in Appendix A and is given in Table 2.1.

Table 2.1 Simple random sample of 100 selling prices (R1 000) of residences sold in Johannesburg from October to December 1991

150.0	107.5	135.0	255.0	268.0	185.0	360.0	240.0	100.0	84.0
155.0	172.6	125.0	97.9	210.0	160.0	360.0	108.0	155.0	255.0
245.0	122.0	300.0	325.0	455.0	170.0	250.0	115.0	70.0	438.0
210.0	135.0	65.0	103.0	172.0	150.6	85.0	158.0	150.0	225.0
85.0	80.0	210.1	152.0	80.0	133.0	95.0	105.0	100.0	76.0
117.0	30.6	69.0	95.0	396.0	159.0	217.5	155.0	135.0	108.0
162.0	110.0	105.0	240.0	60.0	124.0	200.0	230.0	70.0	125.0
138.0	235.0	250.0	93.0	158.0	104.0	105.0	140.0	145.0	70.0
100.0	122.5	170.0	190.0	77.0	186.0	110.0	140.0	90.0	200.0
131.0	132.0	130.0	123.0	175.0	220.0	90.0	45.5	100.0	236.0

Simple random sampling is not always the best method of sampling, but it forms an integral part of all the other probability procedures that will be discussed. For this reason it is important to fully understand the principles underlying this method.

There are various reasons why simple random sampling is not always the best method. The population is sometimes so large that allocation of numbers would take too long, while often there are no recent lists from which all the population elements can be identified. In projects in which the study involves a large geographical area, elements in remote areas might be included in the sample because of random selection. Inclusion of these elements leads to needlessly high costs. Additional information on the population is often available, for instance, that the population elements differ widely with respect to some of the characteristics that are relevant to the survey. A simple random sample obtained from such a heterogeneous population can result in a substantial sampling error. However, it is important to note that, although this method of sampling can result in a substantial sampling error, the sample will not be marred by sampling bias, since each element in the population stands an equal chance of being included in the sample.

Note

If a random sample has to be taken from the population of 376 residences sold in Durban, it is clear that the random numbers 377 to 999 (as well as 000) must be ignored. By associating more than one random number with each population number, random numbers can be better utilized and time and effort can be saved. The following association can, for example, be made:

Random numbers	Residence number
1, 501	1
2, 502	2
.	.
.	.
.	.
376, 876	376

If, for example, random number 683 is drawn, it refers to residence 183. Only the random numbers 377 to 500 and 877 to 999 (as well as 000) are now ignored.

Alternatively, all the random numbers from 001 up to and including 752 (twice 376) could be used by subtracting 376 from all numbers greater than 376 but smaller than or equal to 752 so that a number smaller than or equal to 376 is obtained. Suppose, for example, the following random numbers are obtained:

Random number	Calculation	Residence number
362	362 – 0	362
592	592 – 376	216
915	915 ≥ 752	none

It is clear that different methods can be used to make optimum use of random numbers. However, the method used must not lead to sampling bias. If the random numbers 753, 754, ..., 999 in the above example were to be decreased by 2×376 every time and the corresponding residences with numbers 1, 2, ..., 247 included in the sample, sampling bias would result, because these 247 residences would then stand a greater chance of being included.

4.1.2 Stratified random sampling

Data Set 2 in Appendix A reflects the results of an investigation into the knowledge about and the use of fresh and frozen vegetables in South African households. We can assume with a reasonable degree of certainty that the use of either fresh or frozen vegetables will be influenced by the marital status of the respondent. The mother of a family will probably place a higher premium on the regular preparation of fresh vegetables because her family's health is important to her. Single persons will probably use take-away meals more often. Single persons who do prepare vegetables for themselves will probably prefer frozen vegetables because they are easier to prepare in small quantities. If a sample of, say, 50 respondents from the 'population' of 543 in Data Set 2 were drawn to study the use of fresh and frozen vegetables, it would be important to ensure that both married and unmarried respondents are included proportionately in the sample. If a simple random sample of 50 is drawn from the 543 respondents, of which only 77 are single (see Question 7 of the questionnaire with Data Set 2), then too few (or perhaps too many) from a particular group might be included in the sample. In this case the method of **stratified random sampling** is more suitable than simple random sampling. This method involves the following steps:

(a) Divide the population into subgroups or **strata**. Each stratum is homogeneous with respect to the characteristic being studied. In Data Set 2 the married and unmarried respondents can be regarded as two strata of sizes 461 and 77 respectively. (Note that five respondents did not answer question 7.) Multiple strata, for example based on the number of children in the household, can also be used.

(b) Draw a simple random sample from each stratum. The determination of the sample size within each stratum depends on various factors. If the number of sample elements drawn from a stratum is proportional to the stratum size, it is called a **proportionate sample**. If 14.3% (the percentage of single respondents in the 'population') of the stratified proportionate sample consists of single respondents, and 85.7% of married respondents, the sample is a proportionate sample.

EXAMPLE 2.2

The selling prices of 2 059 houses sold in four large cities from October to December 1991 can be regarded as a relatively heterogeneous population (see Appendix A, Data Set 1). Residences in Johannesburg are generally more luxurious than in any of the other cities in the country. In view of the fact that the other cities probably also differ from one another, it seems reasonable to divide the population of 2 059 selling prices into four strata. If a proportionate stratified random sample of 100 selling prices is taken from each population, the relevant subsample sizes are given in Table 2.2.

The sample size for Johannesburg (n_1), for example, is obtained from the relation

$$\frac{n_1}{100} = \frac{843}{2\,059}$$

so that $$n_1 = \frac{843}{2\,059} \times 100$$

$$= 40.9$$

$$= 41 \text{ (approximately)}$$

Table 2.2 Stratum sizes and sample sizes for a proportionate stratified random sample from a population of selling prices of residences in four large cities

Stratum	Stratum size	Sample size
1. Durban	376 (N_1)	18 (n_1)
2. Johannesburg	843 (N_2)	41 (n_2)
3. Pretoria	434 (N_3)	21 (n_3)
4. Cape Town	406 (N_4)	20 (n_4)
	2 059 (N)	100 (n)

Once the number of elements to be drawn from each stratum has been determined, the sampling is carried out as follows:

Begin, for example, on the fourth page of Table B.1 (Appendix B), in row 31, column 5. As in Example 2.1, read the random numbers as three-digit numbers, but move vertically down: columns 5, 6 and 7 of row 31 form the first number, columns 5, 6 and 7 of row 32 the second number, and so forth.

In this way 18 three-digit numbers smaller than or equal to 376, 41 three-digit numbers smaller than or equal to 843, 21 three-digit numbers smaller than or equal to 434 and 20 three-digit numbers smaller than or equal to 406, are drawn.

Durban

88	306	145	261	277	218	40	214	342	295
17	66	234	139	59	52	91	72		

Johannesburg

329	262	677	581	71	16	487	841	164	342
627	117	278	593	622	288	210	755	418	468
275	332	633	635	401	152	18	415	814	620
104	792	125	56	33	304	751	306	716	439
420									

Pretoria

138	139	383	370	10	167	194	315	7	44
97	178	33	52	66	273	262	274	32	260
376									

Cape Town

90	15	362	131	404	204	58	17	279	114
238	387	101	364	303	291	308	187	147	174

The proportionate stratified random sample of 100 selling prices (R1 000) is now drawn from Appendix A, Data Set 1 and reflected in Table 2.3.

Table 2.3 Proportionate stratified random sample of 100 selling prices (R1 000) of residences that were sold in four large cities from October to December 1991

Durban

75.0	135.0	160.0	117.0	79.0	340.0	265.0	85.0	110.0	143.0
40.0	180.0	125.0	147.0	120.0	170.0	130.0	161.5		

Johannesburg

95.0	380.0	190.0	120.0	332.0	212.0	78.0	165.0	58.0	265.0
70.0	185.0	131.1	115.0	165.0	100.0	61.0	170.0	205.0	215.0
172.6	114.7	117.0	370.0	268.0	210.0	118.0	186.0	90.0	105.0
160.0	320.0	335.0	95.0	330.0	70.0	157.3	110.0	210.1	95.0
250.0									

Pretoria

111.0	165.6	90.0	104.5	140.0	251.8	210.0	115.0	70.0	202.9
74.0	108.0	106.0	126.0	270.0	80.0	190.0	228.0	155.0	165.0
185.0									

Cape Town

70.0	87.0	85.0	135.0	240.0	194.0	110.0	83.4	120.0	89.9
320.0	89.0	81.5	117.0	89.0	75.0	130.0	270.0	155.0	325.0

A proportionate stratified random sample is not necessarily the best stratified sample, since the strata can differ considerably with respect to inherent variation. A small number of observations will suffice for representing a homogeneous stratum, while a large number of observations is required from a

stratum that is heterogeneous with respect to the characteristics being studied. There will probably be more different uses for vegetables (see Appendix A, Data Set 2) among the families of married respondents than among the single respondents. It could therefore be meaningful to draw more than a proportionate number of observations from the stratum of married respondents and fewer from the stratum of single persons. For a relatively small sample of 50, however, this could lead to bias, although the sampling error would become smaller.

Another factor to be considered when determining the sample size within each stratum, is the cost involved in drawing the sample within each stratum. In many surveys geographical areas form natural strata, for instance, the different regions in a country or the suburbs of a city. If the centre in which the survey is being done is located in one stratum, the cost involved in the sampling within that stratum would be lower than it would be in strata situated further away. For this reason, relatively more observations can be drawn from where the cost is lower, and fewer observations from the other strata. The following rule was formulated by Cochran for optimum allocation of sample sizes to the strata:

Cochran's rule

Suppose a population of size N consists of p strata of sizes $N_1, N_2, \ldots, N_p$. The measures of relative variation with regard to the characteristic being studied are respectively $v_1, v_2, \ldots, v_p$, while the relative cost of the survey per element for the various strata can be indicated respectively by $k_1, k_2, \ldots, k_p$. The number of elements taken from the ith stratum in a stratified random sample of n elements is given by

$$n_i = \frac{\frac{N_i v_i}{\sqrt{k_i}}}{\sum_{j=1}^{p}\left(\frac{N_j v_j}{\sqrt{k_j}}\right)} \times n, \quad i = 1, 2, \ldots, p$$

EXAMPLE 2.3

Labour unrest has developed among workers in the building industry because of alleged poor wages and employment benefits. In order to prevent strikes that could paralyse the industry, the Builders' Trade Union commissions an independent organization to investigate these allegations in the region where the unrest is worst. Three strata of workplaces are identified, namely urban, peri-urban and rural. For the purpose of this investigation, peri-urban areas are regarded as workplaces outside cities, but within a radius of 50 km of an urban area. Table 2.4 reflects the estimated values of N_j, v_j and k_j for all three strata. These values could have been drawn from the trade union's records and from estimates that were obtained in a pilot study. If, in accordance with Cochran's rule, 300 workers have to be selected from the total number of 85 000 workers by means of a stratified random sample to complete questionnaires, the sample sizes would be as indicated in Table 2.4.

The sample size for the first stratum (urban area), for example, would be calculated as follows:

$$n_1 = \frac{\frac{65\,000 \times 1.0}{\sqrt{5}}}{29\,068.9 + 9\,185.6 + 4\,419.4} \times 300$$

$$= \frac{29\,068.9}{42\,673.9} \times 300$$

$$= 204.36$$

$$= 204 \text{ (approximately)}$$

Table 2.4 Approximate sample sizes for the three strata used in the investigation into labour unrest, as determined by means of Cochran's rule

Stratum	N_j	v_j	k_j	$\frac{N_j v_j}{\sqrt{k_j}}$	n_j (approximately)
1. Urban	65 000	1.0	5	29 068.9	204
2. Peri-urban	15 000	1.5	6	9 185.6	65
3. Rural	5 000	2.5	8	4 419.4	31
	85 000			42 673.9	300

When a stratified random sample has to be drawn from a population without natural strata, it is often difficult to determine how many strata should be created. If it is possible to regard the population as heterogeneous with respect to a few characteristics, stratification can be applied successfully in more than one direction. In a market research survey the population can, for example, be stratified according to gender and socio-economic grouping. If each population element falls within one of three socio-economic groups, the population is then divided into six strata.

The advantages of stratified random sampling over simple random sampling are the following:

(a) It provides a sample with a smaller sampling error and is therefore more reliable. The more strata used, the smaller the sampling error.

(b) It provides additional information on subgroups (namely the strata) and this makes it possible to study the subgroups separately. With multiway stratification, however, care should be taken to ensure that enough elements are drawn from subpopulations to describe these subpopulations adequately.

As with simple random sampling, the greatest disadvantage of this method is that a complete list of all the population elements must be available before it can be applied. In addition, the number of population elements and also a

certain degree of information on the characteristics to be studied in each stratum must be available. The application of both simple random sampling and stratified random sampling, as explained above, is a time-consuming process. If, however, the population elements are captured on computer, these two methods of sampling can be carried out within seconds.

Due to the fact that lists of population elements are not always available and that not everyone who wants to draw a sample has the necessary computer facilities, simpler and quicker probability methods will now be discussed.

4.1.3 Systematic sampling

If a sample of size $n = 10$ has to be drawn from a population of size $N = 200$, it means that one element in every 20 ($\frac{N}{n} = 20$) has to be selected. This principle is used in systematic sampling. The steps in this method involve the following:

(a) Calculate the ratio $\frac{N}{n}$ rounded off to the nearest integer, to obtain the **sample ratio** v.

(b) Arrange the N population elements clockwise in a circle so that numbers 1 and N are adjacent.

(c) Select the first element k in the systematic sample at random from all N elements (use the method described in 4.1.1).

(d) Obtain the other $(n - 1)$ elements by starting with k and moving clockwise around the circle and selecting every vth element until n elements have been selected.

This method is known as the **circular method** of systematic sampling.

EXAMPLE 2.4

Suppose the population being studied is the areas of the buildings of the 406 residences sold in Cape Town from October to November 1991 (see Appendix A, Data Set 1). A systematic sample of size 50 has to be drawn from this population. The sampling ratio is calculated as follows:

$$\frac{N}{n} = \frac{406}{50}$$
$$= 8.12$$

The sampling ratio is therefore 8.

Now arrange the 406 population elements clockwise in a circle as indicated in the schematic representation in Figure 2.1.

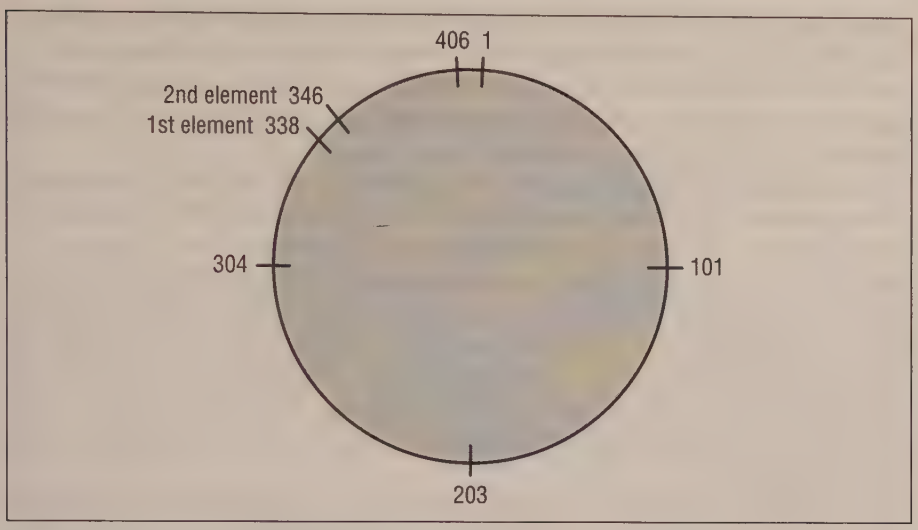


Figure 2.1 Circular method of systematic sampling

Suppose the number k is selected at random as 338. The following element numbers (the first two of which are indicated in Figure 2.1) identify the systematic sample:

338	346	354	362	370	378	386	394	402	4	12
20	28	36	44	52	60	68	76	84	92	100
108	116	124	132	140	148	156	164	172	180	188
196	204	212	220	228	236	244	252	260	268	276
284	292	300	308	316	324					

The systematic sample of the areas of buildings is reflected in Table 2.5.

Table 2.5 Systematic sample of areas (m^2) of 50 residences in Cape Town (see Appendix A, Data Set 1)

152	88	247	143	127	166	146	210	140	80
80	115	217	176	85	120	130	256	210	90
101	124	344	254	150	169	162	263	172	93
170	175	97	149	198	237	220	109	103	114
97	105	103	255	91	109	100	133	90	149

Availability of a complete list of all the population elements is not always a requirement for drawing a systematic sample. Items just completed in a production process form a natural sequence and if quality control is required, systematic sampling is the obvious method. The initial selection of a completed item is done at random and after that every tenth item, for example, is then subjected to inspection. The advantage of this method is that the production process can be interrupted immediately if the quality is no longer up to standard.

A further example of systematic sampling is an opinion poll by the local authority in certain city suburbs to determine what the residents' needs are with

respect to branch libraries, local medical clinics, more extended bus routes, more traffic lights at crossings and so forth. To obtain this information it is decided that visits will be paid to residents in, for example, every sixth house on the eastern side of north-south streets and on the northern side of east-west streets.

Since systematic sampling is easy to apply and very flexible, this method is often preferred to simple random sampling. If the population is relatively homogeneous with respect to the characteristics being studied, a systematic sample is as reliable as a simple random sample. In heterogeneous populations, correctly executed systematic sampling is better than simple random sampling because the nature of systematic sampling ensures that all parts of the population are represented. If Data Set 2 of Appendix A is listed in two groups according to the marital status of the respondents, a systematic sample from this joint list would be equally as representative as a proportionate stratified random sample.

A potential danger in systematic sampling is the phenomenon of **lurking periods**. In the example of the opinion poll conducted by the local authority by calling at every sixth house, problems can develop if all the street blocks in a certain suburb consist of six houses and the initially selected house is then, by chance, a corner house. This will mean that the sample will be confined to the residents of corner houses and although the attitudes of these families with respect to library and medical facilities should be representative of the population, they could provide biased information on traffic problems. If the population characteristic involves a time factor (for example: the daily sales of a shop, or the hourly electricity consumption of a city) and a systematic sample is drawn, care should be exercised to ensure that the sampling ratio does not correspond with the length of a possible period that could be present in the population (for instance 7 in the case of daily sales and 24 in the case of hourly electricity consumption).

If the sampling ratio for drawing a systematic sample is prescribed at, say, 20 (this means that 5% of the population has to be included), the sample can be drawn by taking a random population element from the first 20 and then every 20th element. This method can be used, for example, in a botanical survey in which 5% of the trees in a plantation have to be included in the sample. A systematic sample is particularly suitable in this case because possible heterogeneities in the soil composition that could affect growth would be properly accounted for in such a sample.

4.1.4 Cluster sampling

If no information, other than the broad framework, is available on a population and a relatively representative sample has to be drawn within a short time without incurring high costs, a cluster sample would be the answer. The procedure involves the following:

(a) Divide the population into natural subgroups or **clusters**. If the advantages of cluster sampling are to be effectively utilized, the elements in each cluster should be geographically close to each other. The houses in a street

block or suburb, the employees in one factory and the Standard 10 pupils at a certain high school are examples of clusters (naturally of different populations).

(b) Select a few clusters at random. All the elements of the selected clusters now jointly form the cluster sample. The number of clusters selected is determined by the desired sample size and the size of the individual clusters.

A cluster sample is particularly suitable if the internal composition of each cluster is heterogeneous with respect to the population characteristics to be studied. Each page in Appendix A, Data Set 1 (with a few obvious exceptions) contains 100 selling prices of residences in the four largest cities in South Africa. If the population being studied is only Johannesburg, each of the nine pages reflecting information on Johannesburg houses forms a cluster. Each of these clusters is a miniature population and is almost as heterogeneous as the population itself. A random selection of any page (with the exception of the last page, which reflects information on only 43 houses) and the record of all the selling prices on such a page renders a cluster sample of size 100, which would be as representative as a simple random sample or systematic sample of the same size. Naturally this method can be improved by regarding each page as two clusters of size 50 (namely, the data of the first and last 50 residences on each page), and then, for instance, selecting two of the sixteen (once again excluding the last 43) clusters at random to obtain a cluster sample of size 100. It is also possible to consider groups of ten residences each, to select ten random numbers between 00 and 84 and then, by associating these numbers with the corresponding cluster numbers, to compile a cluster sample of size 100.

EXAMPLE 2.5

Assume that we wish to draw a cluster sample of size 125 from the population of Pretoria houses (Appendix A, Data Set 1). Consider the groupings of ten consecutive houses as the respective clusters. There are 434 houses in the population, which means that there are 43 clusters of size 10 and one of size 4. Number the clusters 00 to 43 and draw 14 numbers between 00 and 43 at random. Assume that we start on the first page of the random digits (Appendix B.1) and use the 6th row, columns 41 and 42, reading down. The random numbers are as follows:

Cluster number	26	04	23	36	40	34	28	00	30	43	21	14	11	15
Number of elements in cluster	10	10	10	10	10	10	10	10	10	4	10	10	10	10

All the elements in the clusters with numbers 26, 4, 23, et cetera, up to cluster number 11 are included in the sample. This gives a total of 124 elements. If the first element in cluster 15 is included, the cluster sample is complete. Note, however, that any element in cluster 43 is more likely to be included than any of the elements in the other clusters. A method for overcoming this problem is shown in Example 2.6.

In cases where the internal composition of clusters is strongly homogeneous and the population reasonably heterogeneous, a cluster sample is not very reliable. Large sampling biases and errors are then possible. The only solution is, where possible, to divide the clusters into smaller clusters and to include as many sample clusters as possible in the sample.

The problem of homogeneous clusters is relevant, since the greatest application of cluster sampling is found in sampling from populations of which the elements are found in a widespread geographical area. The clusters could then, for example, comprise the households in suburbs or street blocks that are usually quite homogeneous, particularly with respect to socio-economic characteristics. A cluster sample from such a geographically distributed population is also known as an **area sample** (or zonal sample) and the advantage of drawing such a sample is that fieldworkers need only be concentrated in a few geographical areas.

Note

If the population consists of clusters with greatly differing sizes (such as the clusters in Example 2.6), clusters with probabilities proportional to the cluster sizes can be selected. Where a list of population elements is available, the sample elements can be obtained by numbering the population from 1 to N and then drawing random numbers smaller than N. The clusters within which the selected elements fall, are included in the sample until the predetermined number of clusters comprising the sample is obtained. Naturally a random number associated with a cluster that has already been selected, should be ignored.

EXAMPLE 2.6

Consider Appendix A, Data Set 3, in which the information on 440 babies is recorded. Assume that a cluster sample of size 70 has to be drawn from this population. The first letter of the variable ID, which corresponds with the first letter of the mother's surname, could be used to form heterogeneous clusters. The information in this data set can be summarized as follows:

Table 2.6 Information on the cluster composition of Appendix A, Data Set 3

Cluster number	1	2	3	4	5	6	7	8	9	10	11
Letter of the alphabet	A	B	C	D	E	F	G	H	J	K	L
Number	6	47	19	25	14	5	16	25	13	17	24
Cumulative number	6	53	72	97	111	116	132	157	170	187	211
Cluster number	12	13	14	15	16	17	18	19	20	21	22
Letter of the alphabet	M	N	O	P	R	S	T	U	V	W	Z
Number	37	12	7	25	18	40	10	1	66	12	1
Cumulative number	248	260	267	292	310	350	360	361	427	439	440

Now draw random numbers between 1 and 440, and include the complete clusters containing the selected elements, until a total of 70 elements has been selected.

Use the random numbers in Table B.1, page 2, row 14, columns 18, 19 and 20, reading down. The elements selected are 022 and 338. The corresponding cluster numbers, relevant letters of the alphabet and cluster sizes are:

Number	2	17
Letter of the alphabet	B	S
Cluster size	47	40

All the patients whose surnames commence with a B will thus be drawn, and also the first 23 patients whose surnames begin with S.

4.1.5 Multistage sampling

The four probability procedures discussed thus far can all be classified as one-stage sampling methods. By repeated use of one of the four methods or by combining two or more of these four methods, a researcher with initiative can design a sampling method that could be suited to any specific situation. Methods developed in this way are called **multistage sampling**. Two examples will suffice.

EXAMPLE 2.7

During the period between 1987 and 1992 the government subsidy per student at universities in the RSA decreased in real terms. In consequence, student fees rose in real terms every year. Many students therefore experienced increasing difficulties in paying for their studies. A representative sample has to be drawn from the students at residential universities in South Africa to determine what type of financing (for example, bursaries, loans, own funds) students use to pay their fees.

In 1992 there were fifteen residential universities in South Africa (Vista University, which offers both contact and distance tuition, is excluded from this study because of its dual character). The number of students (head count) at each university in 1991 is given in Table 2.7.

Because the universities are spread throughout South Africa and each has a unique character, the student composition in terms of language, culture and socio-economic background at each university is unique. The population of 161 692 students can therefore be regarded as comprising five strata. The following sampling procedure is designed:

(a) A proportionate sample is decided on. The number of students at each university to be included in the sample appears in the last column of Table 2.7.

(b) Instruction at the various universities is offered mainly within faculties. Often specific faculties, such as medicine and agriculture, are geographically removed from the main campus. If questionnaires have to be given to individual students for completion, then, from a practical point of view, the cost involved could be minimized if students from all the faculties of the

universities need not be implicated in the survey. So, only two faculties at each university are therefore selected for the survey. There are considerable differences between the fees of the various faculties, but the thirty faculties selected should be fairly representative of all faculties.

(c) Faculties usually keep complete records of students, and the students in the two faculties selected at each university can be regarded as a subpopulation. Alphabetical lists of the names of the students in the two selected faculties at a certain university can be combined and once the sampling ratio has been determined, a systematic sample of the desired size can be drawn. It should be easy to obtain the addresses of the selected students from the faculty records. Alternatively, if personal contact is deemed essential for the investigation, students can be contacted during or after lectures. Their whereabouts can be determined with the aid of the information contained in the faculty records. The above systematic sample should ensure that male and female students, graduate and postgraduate students, and the different study courses are all sufficiently represented.

This two-stage sample utilizes all the good characteristics of stratified, cluster and systematic sampling.

Table 2.7 Head count of students at South African residential universities in 1991. Sample sizes are given in the last column

University	Head count	Sample size
Cape Town	14 268	265
Durban-Westville	8 341	155
Medunsa	2 143	40
Natal	13 406	249
North	12 796	237
Orange Free State	9 395	174
Port Elizabeth	4 929	91
Potchefstroom	8 860	164
Pretoria	22 062	409
Rand Afrikaans	9 349	174
Rhodes	4 100	76
Stellenbosch	14 525	270
Western Cape	13 925	258
Witwatersrand	18 532	344
Zululand	5 061	94
Total	161 692	3 000

Source: Report NATED 02-214(91/08), Department of National Education

EXAMPLE 2.8

The map in Figure 2.2 shows the section of Pretoria which the marketing department of a large chain store considers to be its most important outlet. The store is located in a relatively central position in the area. The manager believes that a few hundred houses in the area should be visited by a marketing research

Figure 2.2 Section of a Pretoria streetmap

organization to assess the standard of living and consumer preferences of the average customer. This survey could lead to large shifts in emphasis regarding the range and quality of products sold by the store. The size of the population concerned, namely the number of residential units in the outlet, is unknown, but this information can be obtained from the municipality. Suppose the population size is estimated at 12 000 residential units, which includes houses and flats. On the

basis of this information the researcher conducting the survey decides to involve about 200 units in the survey. He follows the method below:

(a) The population is divided into geographical clusters. Initially it could seem feasible to use the different suburbs as clusters, but a study of the map reveals that suburb sizes vary considerably. A better geographical identification of 12 clusters is indicated with dotted lines in Figure 2.2. The 12 complete grids on the map are used as the point of departure for forming the clusters. In order to simplify step (c), however, cluster borders are demarcated by streets, rails and non-residential areas so that street blocks are not divided by cluster borders. Since virtually all the clusters formed include sections of more than one suburb, greater heterogeneity within the clusters is ensured.

(b) The market researcher now decides to select four clusters at random. By means of random numbers he selects clusters 1, 5, 8 and 10.

(c) In order to realize the aim of 200 residential units, the researcher decides to visit about fifty houses per cluster. Cluster 1 consists of about forty residential blocks, each comprising twenty to thirty residential units. If the blocks are numbered and, for example, three are selected at random, all the residential units in the blocks selected can be visited. This process is repeated in clusters 5, 8 and 10. The blocks selected in each of the four clusters are shaded in Figure 2.2. In clusters that comprise blocks of largely diverse sizes, the blocks can be grouped together to create residential areas of more or less equal size.

The method discussed above is called **two-stage cluster** or **area sampling** and is used quite frequently in practice. If a relatively large number of clusters is selected in the first stage, sampling errors and biases are usually small. Unfortunately, however, the selection of a large number of clusters in the first stage increases the cost considerably. As a result, a compromise often has to be struck between the cost of the investigation and the accuracy of the information.

Note

In sampling studies like the ones described in Example 2.8 the residents of the home units selected are often out. This problem can be solved in different ways. Revisiting the units until the residents are found at home is the best solution, but is impractical and expensive. Alternatively, more than the required number of elements can be drawn from the strata or clusters during the planning of the sample to make provision for non-respondents. If a fieldworker then finds no-one at home, he simply ignores the unit. This procedure is known as **oversampling**.

4.2 Non-probability procedures

Three non-probability procedures will be discussed briefly in this section, namely convenience sampling, judgmental sampling and quota sampling. The former two have no statistical basis and are subjective. Quota sampling, however, is used more often and is therefore discussed in more detail.

4.2.1 Convenience sampling

As the name denotes, this method consists of the population elements that the researcher can study with the greatest convenience. Drawing such a sample is cheap and it can be quickly executed, but it naturally leads to extensive sampling error and bias. The quick interviews that newspapers conduct with a few passersby in a street to obtain the views of the so-called 'man in the street' on a topical subject is a good example of convenience sampling. A further example is a sample consisting of volunteers.

Although the convenience sampling method is definitely not recommended for a project where the results are to be used in any decision-making process, it can be used in pilot studies to identify, for instance, weaknesses in a questionnaire (see Section 6).

4.2.2 Judgmental sampling

This method is used in situations where small samples have to be drawn from heterogeneous populations. The selection of sample elements is left to the 'good judgment' of a person who is familiar with the relevant population characteristics. This person selects elements that, in his opinion, are representative of the population.

Although this method is better than random selection of elements by a layman, it remains subjective and is subject to considerable sampling error and bias. If, for example, the researcher is interested in the success of the corn harvest in a certain agricultural area, the local agricultural extension officers can be asked to designate a few representative farms in the district. These farms form a judgmental sample, and on the basis of their corn harvest, the researcher can get some idea of the success of the corn harvest in the region concerned.

4.2.3 Quota sampling

Quota sampling can be regarded as a simplified form of stratified random sampling. The method was also designed with a view to drawing samples from a heterogeneous population, but in contrast to stratified random sampling, a list of the elements in the population is not a prerequisite. The method amounts to the formation of reasonably homogeneous subpopulations or **cells** by using so-called **control characteristics** for which census figures of the population are available. The sample sizes or quotas to be taken from each cell are usually selected in proportion to the cell sizes, and sample elements are then collected at random by the fieldworkers in the various cells until the desired quotas have been obtained. In an investigation into the earnings of accountants, age can be used as a control characteristic, since the relevant census information is available. Once the accountants have been divided into a few age groups (cells), the quotas can be determined in proportion to the cell sizes and accountants from the different age groups can be included in the sample until the various quotas have been filled.

Notes

(a) Although a quota sample is more reliable than a convenience or judgmental sample, it has no guaranteed good qualities. Considerable sampling errors and bias can therefore be expected. In contrast to probability procedures, there is no method by which the extent of sampling error and bias associated with non-probability procedures can be determined. **Validity studies** are often used to evaluate non-probability procedures. This technique compares sample and census information with respect to characteristics of the population (indicators) for which census information is in fact available. Characteristics such as language, age, gender or religious denomination are often used as indicators. However, validity studies can create false confidence in an inherently poor sample and should be used circumspectly.

(b) The researcher is often confronted by sample information that has been collected with the aid of non-probability procedures. In addition, he might not know what method has been used to collect the information. Is this information totally useless? Although no generalizations can be made with respect to the population from which the sample has been drawn, the information can be interpreted as case studies that describe the behaviour patterns of a section of the population. Such sample information can be useful in planning follow-up or related investigations.

(c) This section has been an attempt to offer the reader a brief overview of the most important sampling techniques used in practice. With this background knowledge the reader should be able to design meaningful sample schemes for different types of research surveys. However, no sampling procedure can be implemented routinely without a careful study of the background and objectives of the investigation concerned.

5 Single sampling plans, multiple sampling plans and sequential sampling

Thus far an important aspect of sampling, namely the determination of the sample size in a particular investigation, has not been accorded much attention. Since this topic is fairly complicated and requires a more advanced background knowledge of statistics, only a single facet of sample size determination will be discussed here, namely single and multiple sampling plans.

In a **single sampling plan** the sample size is determined before the sample is drawn.

In all the examples discussed thus far in this chapter, the sample size is regarded as a given fact that had been determined beforehand, and we can therefore refer to single sampling plans in the examples. It does sometimes happen that after the sample has been drawn, the sample size is recognized as having been unnecessarily large, with subsequent high costs and loss of time. For this reason, **acceptance sampling**, i.e. sampling from a consignment of products offered for sale or delivered on order, often uses double and multiple sampling.

With a **double sampling plan** a sample of size n_1 is first drawn, and on the basis of this sample a decision is made as to whether the consignment is to be accepted, rejected or subjected to further testing. If further testing is required, a second sample of size n_2 is drawn. On the basis of the joint sample of size $n_1 + n_2$ a decision is then made as to whether or not the consignment will be accepted.

A double sampling plan is particularly effective if the consignment is of an exceptionally good or an exceptionally poor quality, since the final decision on such a consignment can then be made on the basis of the relatively small first sample. **Multiple sampling plans** involve a direct extension of the principle of double sampling plans. Example 2.9 explains the matter further.

EXAMPLE 2.9

Suppose a dealer in electrical appliances has to decide about the acceptance or rejection of a large consignment of light bulbs delivered by a certain supplier. This decision can be made only after a sample of light bulbs has been tested to ensure the required quality. For various reasons, including the possible destruction of the light bulbs during testing, the testing should be confined to a sample of the minimum size. Suppose the dealer decides on an initial sample of size $n_1 = 40$. Suppose also that the consignment is rejected if the number of defective light bulbs in the sample is equal to at least 4, the so-called rejection number, while the consignment is accepted if the number of defective light bulbs is equal to a maximum of 1, the acceptance number. If two or three defective light bulbs are found in the first sample, a second sample of size $n_2 = 100$ is drawn. On the basis of the joint sample of size $n = 140$ a decision is then made to accept the consignment if the total number of defective bulbs is a maximum of seven, or to reject the consignment if the total number of defective bulbs is at least eight. This double sampling plan is represented schematically in Table 2.8, and a possible extension to a multiple sampling plan is given in Table 2.9.

Table 2.8 Double sampling plan for acceptance sampling of a consignment of light bulbs

Sample number	Sample size	Combined sample		
		Size	Acceptance number	Rejection number
1	40	40	1	4
2	100	140	7	8

Table 2.9 Multiple sampling plan for acceptance sampling of a consignment of light bulbs

Sample number	Sample size	Combined sample		
		Size	Acceptance number	Rejection number
1	30	30	0	4
2	30	60	2	5
3	30	90	3	5
4	30	120	5	6

When expensive equipment (such as guided missiles) is tested, or when the testing of each sample element takes a long time, multiple sampling can be refined to samples of size 1. This leads to so-called sequential sampling.

With **sequential sampling** observations are taken one by one and after each draw one of the following decisions: 'accept consignment', 'reject consignment' or 'make a further sample observation' is made.

It should be emphasized that the multistage and sequential sampling plans are not designed arbitrarily, but that more advanced statistical techniques are used. A fuller exposition on the determination of sample size in single sampling plans is given in Chapter 11.

6 Designing questionnaires

Once a sampling method has been selected, the physical collection of the data commences. This stage of the investigation is called the **fieldwork** and the persons collecting the data are called **fieldworkers**. Where comprehensive information on each selected sample element has to be obtained, this stage of the research process usually takes the longest.

In a large percentage of sampling surveys the sample elements are persons or families. This tends to be the case particularly in opinion polls and market

research. If the information required from each respondent is extensive, a well-planned questionnaire is indispensable.

When a questionnaire is designed, the following three aspects are extremely important:

(a) obtaining the information from the population as accurately as possible;

(b) obtaining the maximum co-operation from the respondents in the target population; and

(c) facilitating the collection and analysis of the data.

With these factors in mind, it is important to pay attention to content and phrasing of the questions, planning of the types of questions, the appearance and layout of the questionnaire, the type of interview that is to be conducted, organization of the fieldwork and editing and encoding the data that has been collected. Each of these aspects will now be discussed briefly.

6.1 The content and phrasing of questions

Only clear, simple questions that will yield the most essential information should be included in the questionnaire. Make sure that the respondents will in fact be able to provide the required information without having to either conduct a study themselves or do a number of calculations. Questions must not suggest a specific answer. Examples of questions that do not meet these requirements, are:

- "Do you often buy Milk Powder X?" (Does 'often' mean weekly, monthly or quarterly?)
- "Most people believe that price control on agricultural products should be abolished. What do you think?" (The respondent will tend to support the abolishment of price control, since most people support it anyway. The respondent is therefore not forced to reflect carefully on the matter.)
- "Why do you eat frozen vegetables?" (Not clear what information is required and divergent answers such as "Because it is healthy" and "Because it is easier to prepare than fresh vegetables" could be elicited.)

Sensitive questions, for instance on a respondent's criminal record or the use of drugs, can elicit untruthful or inaccurate answers in the case of personal interviews. However, there are methods that can be followed to obtain accurate answers to such questions. These methods will not be elaborated on here.

6.2 Different types of questions

A question that provides two or more answers from which the respondent has to select one (or sometimes more) or which provides extreme values between which the respondent has to exercise a value judgement on the basis of a certain scale, is known as a **closed question**. Different types of closed questions can be distinguished, but only the most important ones will be discussed on the basis of examples:

Dual question:

Gender

Male	
Female	

Multiple choice question with only one response:

Number of children

1	
2	
3	
4 and more	

Multiple choice question with one or more responses:

What type of transport do you use to get to work?

Car	
Bus	
Train	
Walk	
Other	

Arrangement of responses:

Arrange the following characteristics of a car in order of importance to you. Indicate a 1 next to the most important, a 2 next to the second most important, etc.

Characteristic	**Ranking**
Petrol consumption	
Power	
Acceleration	
Reliability	
Price	

Grids:

In your opinion, what type of beverage is preferred by the following groups of consumers?

Type of consumer	**Type of beverage**				
	Fruit juice	Beer	Wine	Fortified wine	Spirits
Teenagers Women Men Sportsmen Sportswomen					

Scale of preference:
What is the probability of your buying a new car within the next year?

Will definitely buy	Strong possibility of buying	Will possibly not buy	Slim chance of buying	Will definitely not buy

Semantic differential scale:
This type of question is an extension of the preference scale. Examples of these are Questions 1 and 6 in the questionnaire used to obtain Data Set 2 in Appendix A.

Open question:
If all possible answers to a certain questionnaire are not known or comprise an extensive collection, an **open question** can be asked. The following question is an example of an open question:

If you live in a flat, why do you prefer a flat to other types of accommodation?

_____

_____

The advantage of an open question such as the above one is that it compels the respondent to reflect carefully before answering. With closed questions one of the possibilities is often selected rashly just to answer the question. However, the editing and statistical analysis of open questions are difficult.

Open and closed questions can also be combined. An example is:

What is your home language?

Afrikaans	
English	
Other (specify)	

6.3 The appearance and layout of the questionnaire

A questionnaire must look good so that a respondent will want to complete it. For this reason, it should be printed on a good quality paper, the printing should be easy to read and the spacing should be such that it is easy to complete. If accurate and reliable responses are to be obtained, the respondent should be able to understand the questions clearly. It is therefore essential, where possible, to provide questionnaires in the mother-tongue of respondents. If a questionnaire has been originally compiled in, say, Afrikaans, and then translated into English or any of the African languages, the translations must be accurate and elegant. Poor language could cause irritability in respondents and this would be detrimental to the investigation. The skilful use of coloured paper (be it one or several colours) also enhances the appearance of the questionnaire. Questionnaires in different colours can also be used to identify strata in the population and to facilitate coding.

The sequence of questions in a questionnaire is extremely important. The first questions should be interesting to catch the respondent's attention. The questionnaire should also be short and concise; ideally not more than two pages, so that it can be printed on a single page. If it is longer than two pages, it should be printed as a booklet rather than on loose pages that are stapled together. In long questionnaires the most important information should be obtained as early as possible since respondents' concentration and interest usually decline as they approach the end of the questionnaire.

6.4 Type of interview

The best questionnaire information is obtained if the fieldworker completes it personally in the presence of the respondent. Since this method is time consuming, the fieldworker can try to collect a group of respondents to complete the questionnaire in his presence. With either of these methods, the fieldworker can explain the meaning of questions and ensure that all the information is obtained from the respondents. To save time and costs, questionnaire information can also be obtained by telephone. Only simple questions (for example dual or multiple choice) can be used in telephone interviews. The disadvantage of this method is that the respondent can terminate the interview at any stage.

The easiest method of obtaining questionnaire information is definitely by sending out postal questionnaires to the persons identified in a sample. Unfortunately, few people respond to postal questionnaires and the information given is often incorrect, inaccurate or incomplete. As already said, this method can lead to biased information due to personality differences between respondents and non-respondents.

If personal interviews are conducted, fieldworkers with pleasant personalities should be chosen. The ability to convince the respondent of the importance of the interview is an essential trait in a fieldworker. Without being insensitive, he/she should be able to keep a foot in the door if necessary. The actual interview should be conducted at a time that suits the respondent and it should be interesting throughout. If the sample is large, it is better to employ a number of fieldworkers rather than to use only one extremely bored and uninspired person.

6.5 Organizing the fieldwork

Appointing reliable fieldworkers is a headache for many researchers. Fieldwork provides students with a good additional income and housewives employed on a temporary basis make reliable fieldworkers with a considerable knowledge of human nature. All fieldworkers should be briefed beforehand on the aim and method of the investigation. Effective control over the fieldworkers is essential. To ensure that visits are made, selected respondents can be called to ascertain whether the visit did take place and what the behaviour of the fieldworker concerned was like.

6.6 Editing and coding the data

Once all the data has been collected, it should be edited critically in order to eliminate any obvious errors. It is essential to determine the reliability of each respondent. This can be done by including a control question in the questionnaire. A **control question** is a hidden repetition of a question that has already been asked in the questionnaire.

Before a questionnaire is finalized, it should be subjected to a **pilot study**. This involves its completion by a few individuals selected arbitrarily from the population. A pilot study identifies ambiguities in the questionnaire and often helps to determine categories that can be used in the closed questions. This facilitates coding and editing.

Despite the above methods to ensure the accuracy of the information collected, there will still be errors in the data. Critical editing of each completed questionnaire is therefore required. Questionnaires that are incomplete or unreliable have to be eliminated. Often the volume of the data collected dictates analysis by computer. If this is the case, the coding of a questionnaire should be discussed in advance with a computer expert. If the responses to each question in a questionnaire are represented by codes, it is easy to capture these codes on computer and then analyse them with the aid of various statistical programmes. In the questionnaire in Data Set 2 in Appendix A the codes are indicated in the right-hand column of the two pages of the questionnaire.

Exercises

1 Consider the population of valuations (Appendix A, Data Set 1, column 4) of 843 residences in Johannesburg.
   (a) Draw a simple random sample of size 25 from this population by reading random numbers from row 33, column 19, on the first page of Table B.1 horizontally from left to right. Record both the random numbers and the corresponding sample of land valuations.
   (b) Draw a systematic sample of size 25 from this population by selecting k (the number of the first sample element) as 712. Record the numbers and the corresponding values of the sample elements.

2 Consider the area of buildings (Appendix A, Data Set 1, column 5) of all the 2 059 residences in the four large cities.
   (a) Draw a simple random sample of size 40 from this population by reading the random numbers in row 7, column 9, on the second page

of Table B.1 horizontally from left to right. Record the random numbers and the corresponding sample elements that are selected.

(b) Draw a systematic sample of size 30 from this population by selecting k (the number of the first sample element) as 1 034. Give the numbers and the corresponding sample areas.

(c) Use the method explained in Example 2.2 to draw a proportionate stratified random sample of size 40 from this population. Obtain random numbers by starting in row 12, column 13, on the third page of Table B.1, reading horizontally from left to right. Record the random numbers as well as the eventual subsamples from the three strata.

(d) Suppose that the data of ten residences in Appendix A, Data Set 1 forms groups (numbers 1 to 10, 11 to 20, etc.). By regarding these groups as clusters, draw a cluster sample of size 50 from the population of 2 059 areas. Explain what you did to obtain this sample. Obtain random numbers by starting to read in row 10, column 2 on the second page of Table B.1 horizontally from left to right.

3 A company with its head office in Johannesburg has five branches located throughout South Africa. Because some of its 3 560 employees are threatening to strike due to poor salaries and wages, the board of directors of the company decides to conduct a comprehensive survey to try and assess, among other things, the standard of living and the productivity of the employees. Since this assignment is urgent, the investigating officer decides to draw a stratified random sample of size 100 employees from the five branches. Only survey costs and stratum sizes are known and the supposition is that the relative variation with respect to the characteristics to be studied is the same at all five branches. Use Cochran's rule to determine how many employees at each branch have to be selected. The table below contains the information available.

Stratum	N_i	k_i
Branch 1	1 230	2
Branch 2	821	3
Branch 3	608	3
Branch 4	111	5
Branch 5	390	4
	3 160	

4 Suppose we want to analyse the response with respect to Question 3 in terms of Data Set 2 in Appendix A by drawing a sample of size 30 from the population of 543 responses.

(a) Draw a random sample by starting in row 15, column 5 on the first page of the random numbers in Table B.1 reading horizontally from left to right.
(b) Draw a systematic sample by taking a random selection of k as 112.
(c) Draw a proportionate stratified random sample by stratifying according to family income (Question 11 of the questionnaire). Use the same starting point as in (b) in the table with random numbers. Record the random numbers and the selected responses to Question 3 for each stratum.
(d) Estimate the proportion of respondents who maintain that frozen vegetables are more expensive than fresh vegetables in all three of the samples included in (a), (b) and (c).

5 In order to study the birth mass of babies a sample of size 25 is required from Data Set 3 in Appendix A.
(a) Draw a systematic sample by taking 303 as the random choice of the starting point k.
(b) Draw a proportionate stratified random sample by stratifying according to
    (i) the gender of the baby; and
    (ii) the geographical location of the doctor's practice (last letter of ID variable, viz P = Pretoria and T = Trichardt).
    In both cases, start by using the random numbers on page 3, column 17, row 17 of Table B.1.
(c) Draw a cluster sample in a manner similar to that indicated in Example 2.6, but assume that the random numbers 055 and 358 were selected consecutively.
(d) For each of the above samples, calculate the average birth mass of the 25 babies. Also determine for the samples in (b)(i) and (b)(ii) the average birth mass of the babies in the various strata.

6 Explain, with substantiation, which method of sampling you would use in the following situations:
(a) A large textile firm wishes to determine the extent to which ladies who buy clothes follow the instructions for use as indicated on the labels. The population being studied is all the adult women in Pretoria.
(b) The Society for the Prevention of Cruelty to Animals is concerned about the poor care of pets by Pretorians who go on holiday. A questionnaire is compiled to study pet care practices, but due to limited financial means, only 1 000 families can be involved in the survey.
(c) A new compulsory subject for Standard 6 pupils is introduced by a certain education department. Having offered this subject for a year, the department wishes to determine the pupils' attitude towards the

subject. Due to limited time the department decides to include only 10% of the pupils exposed to the subject in the survey.

(d) The Cape hakea is an invader plant that threatens the survival of the indigenous vegetation on the mountains of the Western and Southern Cape. The Department of Forestry wishes to determine the extent of the infestation (mean number of plants per hectare) of a mountain region of approximately 2 000 square kilometres. Due to many physical restrictions it is possible to study only 200 hectares of this area relatively extensively.

(e) The effect of the 1992 drought on different types of farming in the OFS has to be determined by way of sampling in the course of a week.

(f) Value added tax (VAT) was introduced on 1 October 1991 and replaced General Sales Tax (GST). Now that the new tax system has been in operation for a considerable period, a survey is to be conducted among all residents of the RSA to determine the attitude of the general public to VAT in relation to GST.

(g) Since a dealer's administration of VAT differs considerably from that of GST, a survey among dealers would provide important information on the fairness and the application of VAT as opposed to GST.

(h) In order to determine whether the number of veterinary surgeons currently being trained is sufficient to meet the future needs of the country, the current number of livestock (sheep and cattle separately) and the annual growth in livestock have to be estimated.

(i) On New Year's Day there were approximately 50 000 people on Durban's beaches. In order to determine whether the facilities (for example: toilets, refreshment points, beach patrols) satisfied the visitors' needs, an opinion poll is to be conducted among at least 200 of the visitors.

(j) It is important to limit elephant numbers in a relatively large game reserve to prevent water shortage and extensive damage to the natural vegetation (habitat). As it is impossible to count the elephants the number has to be estimated by means of sampling.

(k) The sports reporter of a newspaper wants to disclose the favourite team on the eve of a rugby test. Since there are long queues of prospective spectators waiting to buy tickets two days before the test, he decides to conduct an opinion poll among them.

(l) A farmer has to examine 30 sheep from a kraal with 583 wool sheep in order to determine the quality of their wool.

(m) A chain store markets a new product in Pretoria. After this product has been on the market for a month, they plan an opinion poll to reflect the public's opinion on the product. At least 200 persons have to be questioned.

(n) To investigate the potential for economic growth in South Africa, a representative sample has to be obtained from every type of company.

(o) A survey is conducted on the use of fax machines by telephone subscribers in Pretoria.
(p) The Pharmaceutical Society of South Africa plans a survey to determine the mean income of pharmacists in South Africa.

7 Compile short questionnaires that could be used in the investigation in 6(a), (b), (c), (g) and (m). The questionnaires should not contain more than five questions.

8 Study the 434 residences in Pretoria in Appendix A, Data Set 1 as a population and assume that the selling price of a house is the characteristic being studied.
(a) Draw a simple random sample of size 30 from this population by regarding the random number on the third page of Table B.1 in row 1, columns 11 to 13 as the first random number and then reading off consecutive random numbers by moving down in the same three columns. Record the random numbers and the sample values obtained.
(b) Since expensive houses are usually built on large stands, we can stratify this population according to the size of the stands. We will then ensure that the different selling prices are adequately represented in a sample. Consider the following stratification:

Stratum 1 – Houses with small stands (size = 1)
Stratum 2 – Houses with medium size stands (size = 2)
Stratum 3 – Houses with large stands (size = 3)

Now obtain a proportionate stratified random sample of size 30 by reading off random numbers as in (a), but use the second page of Table B.1.
(c) Draw a systematic sample of size 30 by starting with house no 89.
(d) Calculate the average selling price per residence for the above three samples. Is there a big difference?

9 With reference to Example 2.3, suppose Table 2.4 is the following:

Stratum	N_i	v_i	k_i
1. Urban	20 500	1	4
2. Peri-urban	10 200	2	8
3. Rural	2 100	3	10

If a stratified random sample of size 80 has to be drawn from the population of workers, use Cochran's rule to determine the sub-sample sizes n_1, n_2 and n_3.

10 A rural constituency comprises seven municipal areas, of which the number of voters per area is indicated below:

Môreskof	–	2 103
Vanwyksvlei	–	1 312
Piesangvallei	–	1 691
Ventershoogte	–	936
Dassieskop	–	1 810
Niekerkspos	–	215
Leeufontein	–	2 936

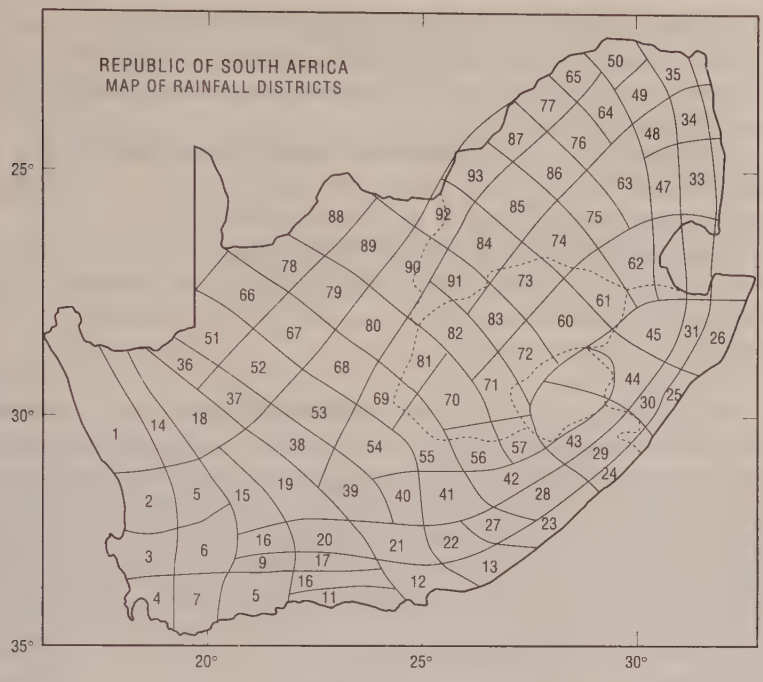
(a) If a proportionate stratified random sample of 200 voters has to be drawn in order to determine the support for a certain candidate in an upcoming election, calculate the sample sizes for each stratum.
(b) Explain how you would go about drawing a two-stage area sample of size 200 from the above population in practice.

11 The following street block is one of the strata of a target population. A fieldworker has to visit seven residences in this stratum in a systematic way. Use the circular method of systematic sampling to select the houses that have to be visited if the fieldworker:
(a) starts with plot number 15;
(b) starts with plot number 28.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
30	29	28	27	26	25	24	23	22	21	20	19	18	17	16

12 Draw (or obtain) a map of a campus on which all buildings, roads, entrances, parking areas, and also the surrounding streets are indicated. In order to study parking problems in and around the campus, a probability sample of 200 students and 50 lecturers must be drawn on the basis of the map.
(a) Explain how you would draw such a sample.
(b) Compile a questionnaire of no more than 10 questions with the aid of which you would obtain the required information on possible parking problems. In your questionnaire include all possible options pertaining to the different types of questions as indicated in paragraph 6.2.

13 The accompanying map indicates the 93 rainfall districts of the RSA. Assume that you have to draw a representative sample of ten districts for meteorological purposes.
(a) Discuss the method you would use. Mention all the factors that you would consider and additional data that you might require in order to implement the best possible method.
(b) Indicate the districts included in your sample.



14 A researcher is interested in investigating the quality of service of supermarkets in Pietermaritzburg. A questionnaire was designed to be distributed among clients of supermarkets to evaluate their experience of service quality in the supermarket. The complete list of supermarkets in the city is supplied with the total monthly turnover of each supermarket.

Supermarket	Monthly turnover R
Spar	**336 000**
City Square	90 000
Commercial Foodliner	43 000
Park Lane	48 000
Kwik Save	55 000
Laager Centre	45 000
Southgate	55 000
Checkers	**180 000**
Cascades	80 000
Scottsville	100 000
Pick & Pay	**191 000**
Hayfields	76 000
Longmarket	115 000
OK Bazaars	**50 000**
Boshoff street	50 000

Data obtained from Steward McArthur, MBL4, 1992

(a) Design a stratified sample of size 500 to determine the number of clients to be interviewed at each supermarket. Use the monthly turnover as the weight for the number of clients to be interviewed at each supermarket.

(b) It is known that one out of three clients is male. How would you divide the sample quota further to ensure that both sexes are represented in the sample?

(c) Which procedure would you follow to conduct the survey? Take into account the day of the week and time of day in your planning.

(d) Discuss the advantages and disadvantages of the sampling procedure proposed in (c).

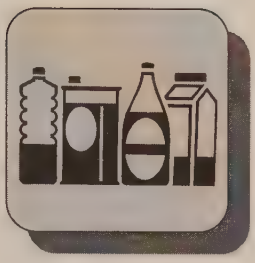
15 In a steel manufacturing company work groups are used to optimize production management. The work groups consist of a supervisor and five to seven employees including a group leader in each group. The management is interested in investigating the functioning of the group by means of interviews with samples of individuals. There are approximately 7 000 group members and the group size varies between six and eight members. The composition of the groups in the different departments is as follows:

Department	Number of groups
Metallurgic	388 groups in 4 sections
Rolling-mills	389 groups in 6 sections
Engineering	299 groups in 5 sections
Diverse	122 groups in 5 sections

It is further known that a list of the names of group leaders sorted according to section is available in each department. Design a suitable sampling procedure of size 700 in such a way that all departments and sections are represented in the sample.

CHAPTER 3

Tabulation and graphic representation



1 Introduction

In Chapter 1 the collection of data was shown to be an essential link in research projects undertaken in various disciplines and fields of application. This chapter deals with a number of standard techniques that can be applied for tabulation and graphic representation of data. It is important that data should be presented in a form that can be easily understood and interpreted by the clients of a firm, the suppliers of a product, colleagues in the same organization and fellow researchers.

The entire process of collecting, processing and describing data is usually motivated by a desire to know more about the underlying population. The selection and application of specific descriptive methods often depend on the nature of the data. Numerical data can be classified as **cross-section data** or **time series data**. Time series (or cross-section) data is comprised of observations collected at a specific point in time, for example the turnover in R-million of a sample of 150 companies during a specific year, or the quantity of chrome plated onto steel plates during a production run. A time series consists of observations collected in consecutive time units. The annual turnover of ISCOR since 1958, the monthly production of fuel from coal by SASOL and the size of the RSA's population since 1900 as determined by census surveys, are all examples of time series data.

When analysing time series we are particularly interested in what happens to the variable being observed over a particular period, in order to predict the future behaviour of this variable. Time series data is obtained by means of observation or measurement and is usually recorded in the sequence in which it has been collected. To facilitate analysis and interpretation, the data should be reorganized in such a way that it can be represented by means of a table or graph. This chapter is confined to the processing of cross-section data. The analysis of time series data is discussed in Chapter 7.

2 Arrays and frequency distributions

Without some or other form of systematization, a raw data set comprising a large number of figures has little meaning. The most basic method of systematizing data is to make use of a data array.

A data **array** is obtained by arranging the observations in sequence from the smallest to the largest.

EXAMPLE 3.1

At the steel manufacturing plant of a company steel plating is chromed on both sides. The quantity of chrome on each steel plate is measured to test whether it complies with the client's requirements. Observations of the quantity of chrome (mg/m^2) on 30 steel plates are given in Table 3.1 and serve as an example of statistical data in unprocessed or raw form.

Since the data is arranged in the sequence in which it has been collected, it is difficult to draw immediate valid conclusions with regard to the raw data set. If Table 3.1 is studied observation by observation, it can be determined that the smallest quantity of chrome is 55.3 mg/m^2, while the largest quantity of chrome is 80.6 mg/m^2. It is also possible to obtain an 'average' value by adding up all the individual figures and dividing the total by the number of observations in the data set, namely 30. Apart from this, limited information can be obtained from the raw data. Had Table 3.1 contained more observations, it would have been even more difficult to draw any conclusions.

Table 3.1 A simple random sample of 30 chrome quantities (mg/m^2) as measured on the upper side of steel plating

64.4	60.4	65.6	59.3	80.6	62.0
68.7	67.8	69.7	64.6	68.4	69.0
71.6	59.4	69.6	73.7	74.8	70.2
73.5	60.1	66.9	65.3	63.4	55.3
68.7	77.8	66.5	74.1	68.4	72.3

Data obtained from MBL group PTA 01192

However, once the data has been arranged in ascending order, as in Table 3.2, it becomes easier to draw conclusions from the data and to divide it into categories.

Table 3.2 Array of a sample of 30 chrome quantities in mg/m^2 as measured on the upper side of steel plating

55.3	59.3	59.4	60.1	60.4	62.0
63.4	64.4	64.6	65.3	65.6	66.5
66.9	67.8	68.4	68.4	68.7	68.7
69.0	69.6	69.7	70.2	71.6	72.3
73.5	73.7	74.1	74.8	77.8	80.6

From this array more information can be obtained about the variable being studied than from the raw data. For example, the following is immediately clear:

(a) The chrome quantities vary from 55.3 to 80.6 mg/m^2.
(b) There is a concentration of quantities with values between 62 and 72 mg/m^2.
(c) Approximately half the quantities are 68.4 mg/m^2 and less.
(d) Relatively few quantities are less than 60 or higher than 75 mg/m^2.

A data array is usually easy to use for reorganizing raw data, provided there are not too many observations involved. For large data sets the process of arranging data from small to large takes too long. In addition, the conclusions that can be drawn from an array are relatively limited. If a data set consists of a large number of observations, the data should preferably be presented in a condensed form, but in such a way that there is a minimal loss of original information. This can be done by grouping the data into classes in order to obtain a frequency distribution.

A **frequency distribution** is a table with classes of values and accompanying frequencies.

It is very important that the classes be continuous and that an observation falls within only one specific class. A frequency distribution of the 30 chrome quantities in Table 3.1 is given in Table 3.3.

Table 3.3 Frequency distribution of chrome quantities on 30 steel plates

Chrome quantity mg/m^2	Class mark x	Score	Frequency (Number of plates) f
54.0 and less than 58.0	56.0	I	1
58.0 and less than 62.0	60.0	IIII	4
62.0 and less than 66.0	64.0	𝍸 I	6
66.0 and less than 70.0	68.0	𝍸 𝍸	10
70.0 and less than 74.0	72.0	𝍸	5
74.0 and less than 78.0	76.0	III	3
78.0 and less than 82.0	80.0	I	1
			30

Note

Column 3 of Table 3.3 illustrates the traditional method for counting the values that fall in a specific class. Each observation from the raw data set that corresponds with a specific class is counted by means of a vertical line. Five observations are indicated by drawing the fifth line horizontally through the previous four vertical lines. Quantities of 4, 5 and 7 are represented by IIII, 𝍸 and 𝍸 II respectively.

3 Frequency distributions of continuous data

Before a frequency distribution can be compiled, the number of classes to be used should be determined. If too few classes are used, a great deal of information is lost, while too many classes do not sufficiently condense the data to be of any practical use. The number of classes selected depends mainly on the size and area of variation of the data set. If a frequency distribution is to be used in further numerical processing, selection of a meaningful number of classes is important.

A number of guidelines for determining the number of classes and the width of each class interval are given below:

(a) Five to fifteen classes are commonly used. A useful guideline for determining the number of classes is the following:

Sturge's rule: k, the number of classes in a frequency table, is determined by:

$$k = 1 + 1.4 \ln n$$

where n is the number of observations in the data set and k is rounded off to an integer.

Table 3.4 can be used as a guideline for determining the number of classes in a frequency table for different sample sizes.

Table 3.4 Approximate number of classes in a frequency table for a given sample size n

Number of observations n	Number of classes k
fewer than 30	4 – 5
30 – 50	5 – 6
50 – 100	6 – 7
100 – 200	7 – 9
200 – 500	9 – 10
500 – 1 000	10 – 11
1 000 – 20 000	11 – 15

From Table 3.4 it is clear that the rule regarding the number of classes is reasonably flexible. The following points are also important in the selection of the number of classes and class widths.

(b) The class widths should where possible be the same, for example, a class width of 4 mg/m^2 in the frequency distribution given in Table 3.3. Sometimes the first or last class interval has to be selected as an **open class interval**. This is done when there are a few very small and/or large values in the data set. Such values are called **outliers**. The first two classes in Table 3.3 could, for example, have been replaced by one class, namely 'less than 62.0'. The frequency in this class would then have been 5. An open final class interval at the top could have been '78.0 and more' if, for example, the two highest values had been 88.3 and 89.2.

(c) Select class boundaries and class marks that contribute towards simplifying the compilation of the frequency table without too much loss of information. The class boundaries should be selected in such a way that the class marks are convenient numbers that are easy to use in further calculations. If the first class in Table 3.3 were '54.0 and less than 58.5', for example, the class mark would be equal to:

$$\frac{54.0 + 58.5}{2} = 56.25 \text{ mg/m}^2$$

The class mark of 56.0 for the first class in Table 3.3 is a more convenient value to work with. It is important to note that it is not possible to calculate class marks for open class intervals. For example, if the final class in a frequency distribution is an open interval, it is sound practice to give the largest observation below the table (for instance in a footnote).

(d) The class boundaries should be determined in such a way that each observation can be grouped into only one class. The following representation of the classes in Table 3.3 is incorrect:

54.0 – 58.0
58.0 – 62.0
62.0 – 66.0
.
.
.
78.0 – 82.0

Here the chrome quantity of 62.0 mg/m^2 (see Table 3.2) can belong to either the second or the third class. However, if the classes are defined as in Table 3.3, it is clear that the value 62.0 falls within the third class. Since the method of writing the intervals as represented in Table 3.3 is cumbersome, the following abbreviated method will be used from now on:

[54.0, 58.0)
[58.0, 62.0)
[62.0, 66.0)
.
.
.
[78.0, 82.0)

Note the interval [58.0, 62.0). A value of 58.0 will be grouped into this class, while a value of 62.0 will be grouped into the following class. The square bracket, [, therefore includes the value, while the round bracket,), excludes the value concerned.

With the aid of the above guidelines we shall now compile a frequency distribution for the data set described in Example 3.2.

EXAMPLE 3.2

Consider the sample of 100 residences sold in the Johannesburg area in the fourth quarter of 1991, as given in Example 2.1. A prospective buyer is interested in the price paid per square metre (R/m^2) for the residences. An array of the amounts is given in Table 3.5.

Table 3.5 Array of the price per square metre for 100 randomly selected residences in the Johannesburg area

273.21	275.76	335.05	394.64	397.73
413.79	434.78	469.39	478.26	484.25
514.71	518.52	526.32	534.59	537.63
538.46	546.87	554.01	558.82	566.04
576.92	600.00	602.41	639.27	666.67
668.32	678.83	684.36	712.07	719.70
728.35	731.25	735.29	743.80	745.61
747.83	758.62	770.37	776.70	789.47
794.39	795.93	796.41	800.00	809.35
814.43	846.99	864.86	874.17	877.86
882.35	894.12	894.74	905.41	906.43
911.76	920.12	947.87	964.91	967.74
978.26	979.73	984.46	984.96	1 000.00
1 000.00	1 011.56	1 012.88	1 021.98	1 068.38
1 078.13	1 089.74	1 111.11	1 117.55	1 185.77
1 201.92	1 201.92	1 206.90	1 226.42	1 227.72
1 235.77	1 245.67	1 270.59	1 276.60	1 304.35
1 311.83	1 335.23	1 343.75	1 348.44	1 363.64
1 380.95	1 419.21	1 420.45	1 487.60	1 516.67
1 568.75	1 577.69	1 613.92	1 659.09	1 663.16

Several frequency distributions will now be compiled for the above data set. In accordance with Sturge's rule the number of classes is approximately

$$k = 1 + 1.4\ \ell n 100 = 7.45$$

Consequently, 7 or 8 is a convenient number of classes, but the number to be used in the end will depend on the selected class width and class limits. Let x_{max} and x_{min} respectively represent the largest and the smallest observations in the data set, i.e. $x_{max} = 1\,663.16$ and $x_{min} = 273.21$. If k classes are used, the class width, c, of each interval should be:

$$c = \frac{x_{max} - x_{min}}{k}$$

For $k = 7$ the width c is therefore at least

$$c = \frac{(1\,663.16 - 273.21)}{7} = 198.56 \approx \text{R}200/\text{m}^2$$

and for $k = 8$ the width c is at least $173.75 \approx \text{R}175/\text{m}^2$.

Wherever possible, the class width and class boundaries should be selected in numbers with which the user can easily identify. The following sets of classes can perhaps be selected for a frequency distribution of the data in Table 3.5.

(a)	(b)	(c)	(d)
$k = 8;\ c = 200$	$k = 8;\ c = 200$	$k = 7;\ c = 200$	$k = 9;\ c = 175$
[100, 300)	[250, 450)	[270, 470)	[250, 425)
[300, 500)	[450, 650)	[470, 670)	[425, 600)
[500, 700)	[650, 850)	[670, 870)	[600, 775)
[700, 900)	[850, 1 050)	[870, 1 070)	[775, 950)
[900, 1 100)	[1 050, 1 250)	[1 070, 1 270)	[950, 1 125)
[1 100, 1 300)	[1 250, 1 450)	[1 270, 1 470)	[1 125, 1 300)
[1 300, 1 500)	[1 450, 1 650)	[1 470, 1 670)	[1 300, 1 475)
[1 500, 1 700)	[1 650, 1 850)		[1 475, 1 650)

The selection of (a) to (c) has the advantage that the class marks are also integers, while this is not the case in (d). The class mark in the first class of (a) is R200/m^2, while the class mark in the first class of (d) is R337.5/m^2.

When choosing classes and class boundaries it is important to remember that the purpose of the frequency distribution is to simplify a data set, and for this reason any of the above class boundaries would be acceptable. However the selection of (a) or (b) would probably be the best. The first class limit need not be the smallest value in the data set, and a smaller number can be used with success to select convenient class boundaries. If the class width c is calculated according to the above formula, the number of classes often works out to be one more than k in practice, particularly if the starting point of the class intervals is considerably lower than the smallest observation x_{min}.

It is often necessary to represent more than one data set in frequency distributions in order to compare the data sets. In such a case it is important to select identical class boundaries for both representations. Since there is considerable freedom in the selection of class boundaries, x_{max} and x_{min} should be taken as the maximum and minimum of all the data sets collectively.

Table 3.7 reflects the frequency distribution of the data in Table 3.5 on the basis of the class boundaries as in (a) above. Tables 3.6 and 3.8 show the influence of the selection of the number of classes, k, and class width, c, on the description of data. In the graphic representation of the data the influence of these two selections will be explained in more detail.

Table 3.6 Frequency distribution of the price per square metre paid for 100 houses in the Johannesburg area, $k = 4;\ c = 400$

Class	Class interval Price per m^2	Class mark x	Frequency f
1	[200, 600)	400	21
2	[600, 1 000)	800	43
3	[1 000, 1 400)	1 200	27
4	[1 400, 1 800)	1 600	9
			100

Table 3.7 Frequency distribution of the price paid per square metre for 100 houses in the Johannesburg area, $k = 8$; $c = 200$

Class	Class interval Price per m²	Class mark x	Frequency f
1	[100, 300)	200	2
2	[300, 500)	400	8
3	[500, 700)	600	18
4	[700, 900)	800	25
5	[900, 1 100)	1 000	19
6	[1 100, 1 300)	1 200	12
7	[1 300, 1 500)	1 400	10
8	[1 500, 1 700)	1 600	6
			100

Table 3.8 Frequency distribution of the price paid per square metre for 100 houses in the Johannesburg area, $k = 15$; $c = 100$

Class	Class interval Price per m²	Class mark x	Frequency f
1	[200, 300)	250	2
2	[300, 400)	350	3
3	[400, 500)	450	5
4	[500, 600)	550	11
5	[600, 700)	650	7
6	[700, 800)	750	15
7	[800, 900)	850	10
8	[900, 1 000)	950	11
9	[1 000, 1 100)	1 050	8
10	[1 100, 1 200)	1 150	3
11	[1 200, 1 300)	1 250	9
12	[1 300, 1 400)	1 350	7
13	[1 400, 1 500)	1 450	3
14	[1 500, 1 600)	1 550	3
15	[1 600, 1 700)	1 650	3
			100

Usually only one frequency distribution is compiled for a data set. Here, more than one was compiled to illustrate the influence of the selection of class boundaries and class widths. In the graphic representation of the data the influence of these selections will be explained in more detail.

4 Graphic representation of continuous data

A frequency distribution is usually represented in two ways, namely with the aid of a **histogram** or a **frequency polygon**. These visual representations enable us to interpret a data set quickly.

4.1 The histogram

The following four points provide guidelines for compiling a histogram:

(a) The class frequencies are plotted on the y-axis and the class intervals according to a convenient scale on the x-axis.
(b) It is common practice to leave spaces that are equal to the class width on the left-hand side of the smallest class boundary and on the right-hand side of the largest class boundary.
(c) With the class intervals as a basis, rectangles are now drawn so that the area of each rectangle is proportional to the corresponding class frequency.
(d) It is important to provide the histogram with a full caption, to indicate clearly what the x- and the y-axes represent and to divide the axis systems clearly.

On the basis of these guidelines, the histograms in Figures 3.1 to 3.3 were compiled for the frequency distributions of Tables 3.6 to 3.8. In Figure 3.1 the second class has the highest frequency. In Figure 3.2 the classes with the highest frequencies are in the middle, while the frequencies gradually decrease to either side of the histogram. However there are a number of extreme values higher than R1 300/m^2, which indicates that the distribution of the average price paid per square metre is **positively skewed**.

We have already mentioned that the selection of the number of classes is reasonably flexible. If too few classes are selected, as in Figure 3.1, too much information will be lost. On the other hand, if too many classes are used as in Figure 3.3, there are irregularities that could be avoided by selecting fewer intervals. If we regard the histogram as an approach to a smooth curve, a comparison of these three figures reveals that Figure 3.2 provides the best representation.

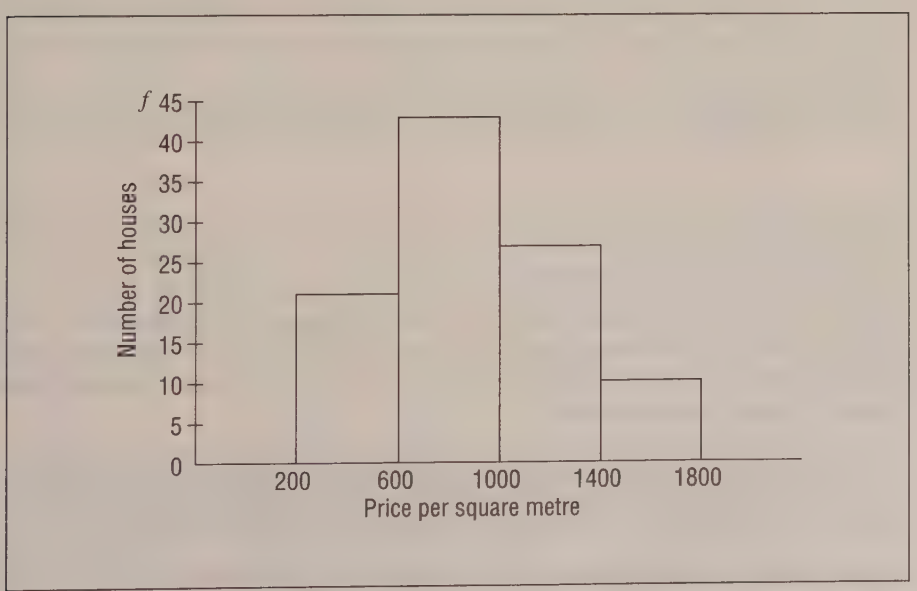


Figure 3.1 Histogram of the price paid per square metre for 100 houses, $k = 4$; $c = 400$

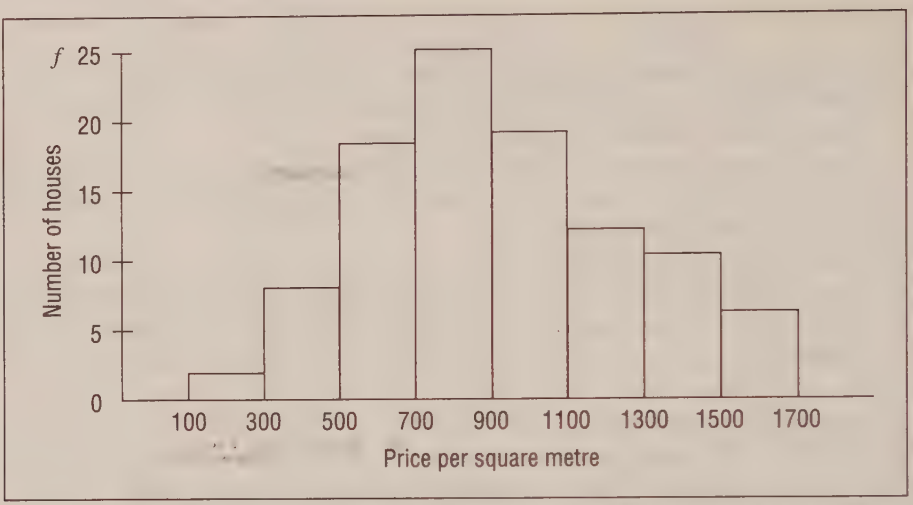


Figure 3.2 Histogram of the price paid per square metre for 100 houses, $k = 8$; $c = 200$

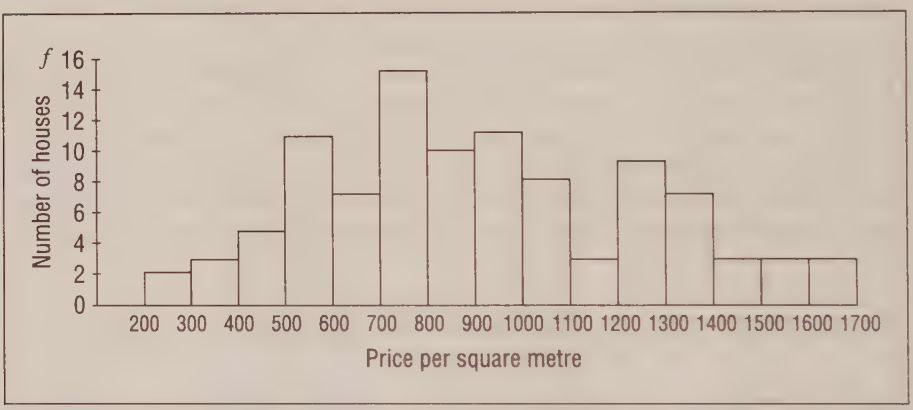


Figure 3.3 Histogram of the price paid per square metre for 100 houses, $k = 15$; $c = 100$

4.2 The frequency polygon

If a histogram of a data set has already been drawn, the frequency polygon can easily be obtained by connecting the top horizontal lines of the rectangles with straight lines. To allow the polygon to connect with the x-axis, additional classes, each with a frequency of zero, are added to the top and bottom of the existing classes. To summarize:

> A **frequency polygon** is a graph in which the points with co-ordinates (class mark, class frequency) are connected by means of straight lines.

Figure 3.4 shows the frequency polygon for the data in Table 3.7. The dotted lines indicate how the polygon can be drawn from Figure 3.2.

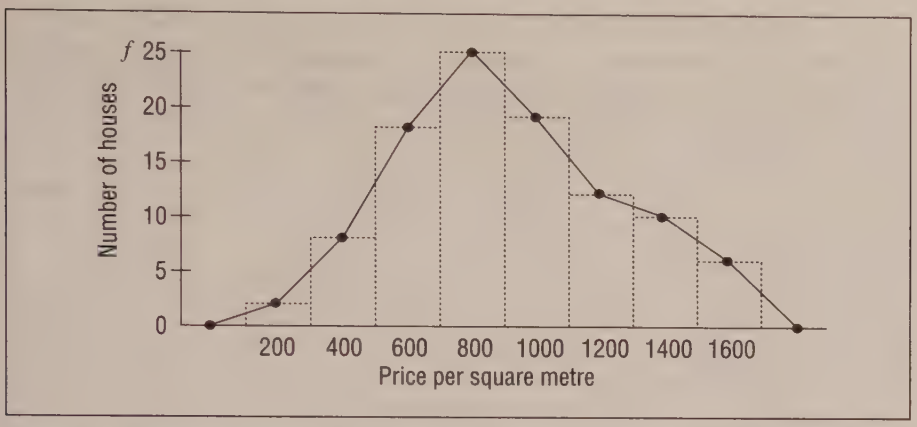


Figure 3.4 Frequency polygon of the price paid per square metre for 100 houses in the Johannesburg area

5 The cumulative frequency distribution and polygon

The **cumulative frequency distribution** is obtained from a frequency distribution by means of the accumulation of the frequencies.

EXAMPLE 3.3

In the construction of the cumulative frequency distribution the cumulative frequencies are compared with the class boundaries and not with the class marks. Table 3.9 shows the cumulative frequency distribution obtained from the frequency distribution of chrome quantities plated to the upper side of steel plate. This cumulative frequency distribution was compiled from the data in Table 3.3. The cumulative frequencies are indicated by F. Note that the cumulative frequency, F, opposite the smallest class boundary is always equal to 0, while F opposite the biggest class boundary is always equal to the sample size n.

Table 3.9 Cumulative frequency distribution of the chrome quantity on 30 steel plates

Class boundary Chrome quantity mg/m² x	Cumulative frequency F
< 54.0	0
< 58.0	1
< 62.0	5
< 66.0	11
< 70.0	21
< 74.0	26
< 78.0	29
< 82.0	30

A graphic representation obtained by connecting the points (x, F) of the cumulative frequency distribution is known as the **cumulative frequency polygon**. This polygon is used mainly for interpolation purposes. Take for example Figure 3.5, which is the cumulative frequency polygon for Table 3.9.

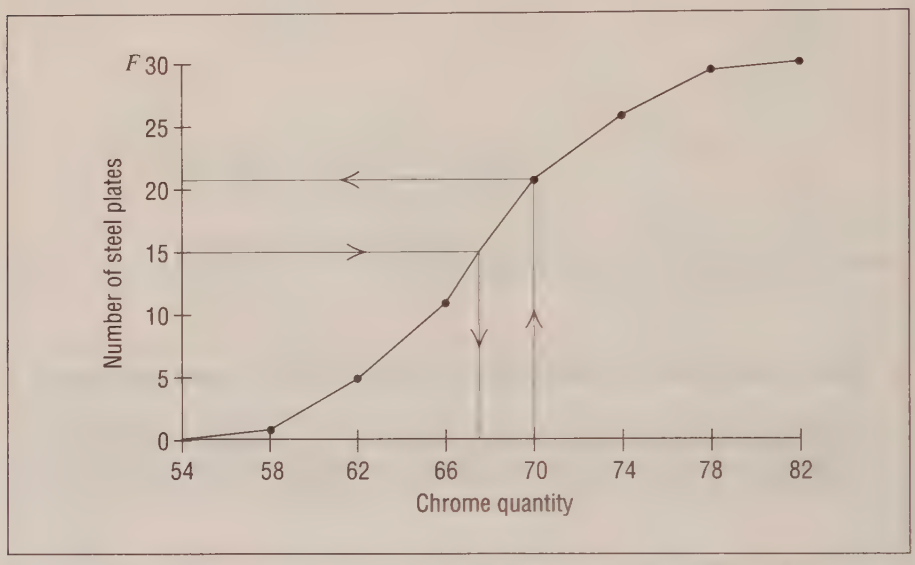


Figure 3.5 Cumulative frequency polygon of the chrome quantity on 30 steel plates

If a chrome quantity of 70 mg/m^2 is selected on the horizontal axis, it can be read from the vertical axis that about 21 out of the 30 chrome quantities are less than 70 mg/m^2. Conversely, a point can be selected on the F-axis and the corresponding value be read on the x-axis. If the value 15 is selected on the F-axis, the figure will reflect that 15 of the 30 chrome measurements on the steel plates, i.e. half, were less than 67.5 mg/m^2.

6 Relative frequency distributions

Relative frequency distributions can be used to compare data sets with differing numbers of observations. If two samples differ radically with regard to size, ordinary histograms or frequency polygons can hardly be used to compare the two data sets. However, if the two frequency distributions have been compiled on the basis of identical class intervals, the relative frequencies can be directly compared.

The **relative frequency distribution** is obtained by expressing the frequencies in a frequency distribution proportionate to the sample size n.

EXAMPLE 3.4

At a manganese processing plant a ferro-manganese alloy is melted down for use in the production of specialized metal alloys. The phosphor content in the alloy has to be constantly monitored to ensure the quality of the product. A phosphor content of more than 0.1% will lower the grading of the alloy, which will mean that the profit margin of the processing plant will drop considerably.

To keep the production costs as low as possible, an alternative supplier of coke is considered. Coke from the alternative supplier is mixed with existing coke supplies to first test the quality of the coke and to decrease the risk of product losses. This coke mixture is called the **experimental coke mixture**. As a control measure the phosphor levels of the existing coke supplies are also monitored and this is called the **control coke mixture**. The phosphor content of the alloy product for the two types of coke mixtures is summarized in Table 3.10.

Table 3.10 Frequency distribution of the phosphor content of manganese alloy for 40 control and 20 experimental coke mixtures

Class interval (percentage phosphor content)	Class mark x	Frequency (control coke mixture) f	Frequency (experimental coke mixture) f
[0.065, 0.070)	0.0675	2	0
[0.070, 0.075)	0.0725	8	1
[0.075, 0.080)	0.0775	12	3
[0.080, 0.085)	0.0825	9	8
[0.085, 0.090)	0.0875	6	5
[0.090, 0.095)	0.0925	3	3
		40	20

Data obtained from MBL group VTR00192

The frequency polygons of the data are given in Figure 3.6. This figure is a misrepresentation of the values in Table 3.10, because there are twice as many observations for the control coke mixture as for the experimental coke mixture.

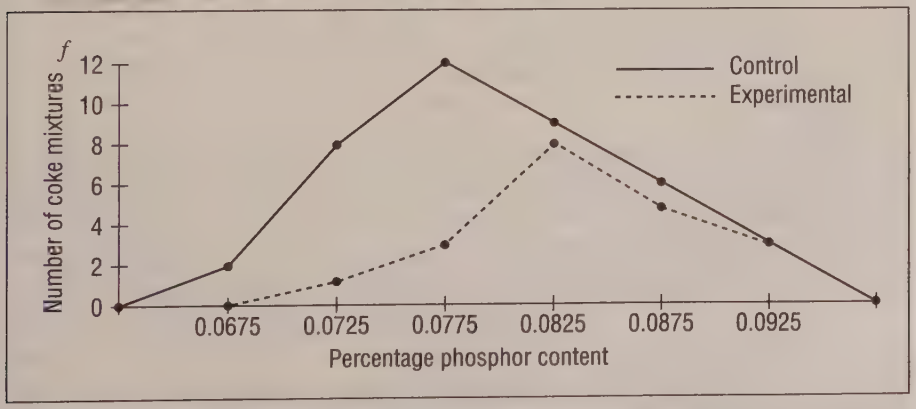


Figure 3.6 Frequency polygons of the phosphor content of manganese alloy for 40 control and 20 experimental coke mixtures

If meaningful comparisons are to be drawn between the distributions of the phosphor contents of the control and experimental coke mixtures, the frequencies have to be expressed in relation to the number of observations. This is necessary because the sample sizes of the two sets of data differ.

The relative frequency distributions of the phosphor content of the experimental and control coke mixtures can be obtained from Table 3.10. Table 3.11 shows this distribution.

Table 3.11 Relative frequency distribution of the phosphor content of manganese alloy for 40 control and 20 experimental coke mixtures

Class interval (percentage phosphor content)	Class mark	Relative frequency f/n (control mixture)	Relative frequency f/n (experimental mixture)
[0.065, 0.070)	0.0675	0.050	0.00
[0.070, 0.075)	0.0725	0.200	0.05
[0.075, 0.080)	0.0775	0.300	0.15
[0.080, 0.085)	0.0825	0.225	0.40
[0.085, 0.090)	0.0875	0.150	0.25
[0.090, 0.095)	0.0925	0.075	0.15
		1.000	1.00

Figure 3.7 shows a comparative frequency polygon of the relative frequency distributions of Table 3.11. This figure enables us to draw a direct comparison between the phosphor content of the test and control mixtures. From the figure we can conclude that the experimental coke mixture possibly renders higher phosphor levels. However, further experimentation and statistical testing (as discussed in Chapters 11 and 12) are required before a final decision can be reached.

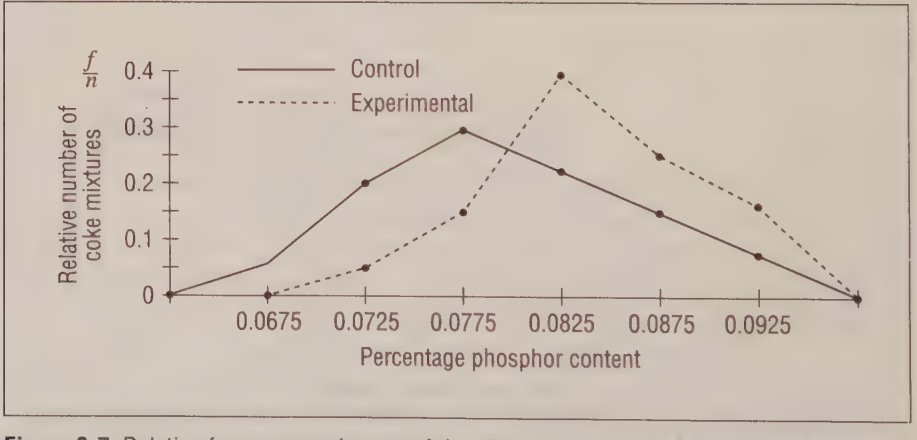


Figure 3.7 Relative frequency polygons of the phosphor content of manganese alloy for 40 control and 20 experimental coke mixtures

From the definition of relative frequencies it follows that the sum of these frequencies should be equal to 1. This provides a measure of control over the calculation of the relative frequencies. An important result is that the relative frequency gives an approximate value of the proportion (percentage) of observations from the population that fall within a specific class. On the basis of Table 3.11, it appears that the chance or probability of the phosphor content in the manganese alloy (melted from the experimental coke mixture) being between 0.080% and 0.085% is about 0.40.

7 The cumulative relative frequency distribution and polygon

The **cumulative relative frequency distribution** is derived from a relative frequency distribution by means of the accumulation of the relative frequencies.

The cumulative relative frequencies for the data in Table 3.11 are calculated in Table 3.12. The cumulative relative frequency polygon for the data in Table 3.12 is given in Figure 3.8. This graph is very useful for interpolation purposes.

Table 3.12 Cumulative relative frequency distribution of the phosphor content of manganese alloy for 40 control and 20 experimental coke mixtures

Class interval (percentage phosphor content)	Cumulative relative frequency F/n (control mixture)	Cumulative relative frequency F/n (experimental mixture)
< 0.065	0.000	0.00
< 0.070	0.050	0.00
< 0.075	0.250	0.05
< 0.080	0.550	0.20
< 0.085	0.775	0.60
< 0.090	0.925	0.85
< 0.095	1.000	1.00

If a phosphor content of, for example, 0.0825 is selected on the horizontal axis, one can read from the vertical axis that about 68% of the measurements of the phosphor content for the control coke mixture were lower than 0.0825, while, for the experimental coke mixture, about 40% were lower than 0.0825. Conversely, a point on the vertical axis can be selected and the corresponding value can be read on the x-axis. If the value of 0.25 or 25% is selected on the F-axis, one can read from the figure that three-quarters of the phosphor level measurements of the control coke mixture were higher than 0.075, while three-quarters of the phosphor level measurements of the experimental coke mixture were higher than 0.081.

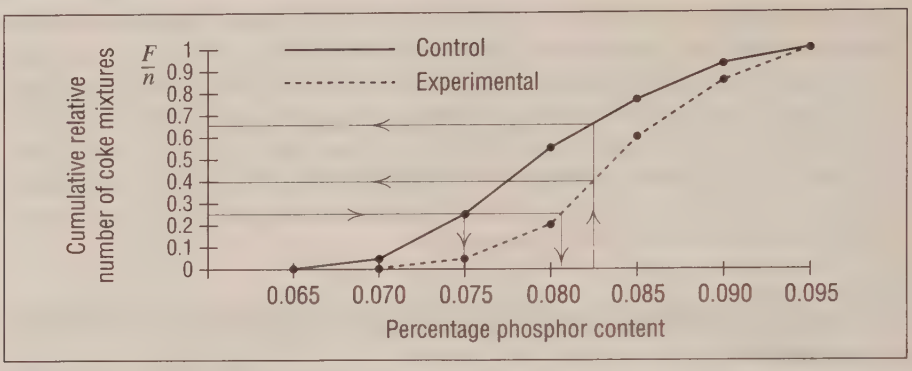


Figure 3.8 Cumulative relative frequency polygons of the phosphor content of manganese alloy for 40 control and 20 experimental coke mixtures

8 Frequency distributions and graphic representations of discrete and categorical data

If the data collected is of a discrete nature, each class in the frequency distribution consists of only a few values and can no longer be represented as continuous intervals, as in the previous frequency tables. Table 3.13 is an example of a frequency distribution for discrete data.

Each 'class' consists of a single figure and this is also the class mark. The range of values that discrete data can assume can be extensive. Examples of these are the number of incoming telephone calls per minute on a switchboard, the number of failures of a transformer in the ESKOM network, and the number of marks obtained by individual candidates in a specific examination. In cases where many values appear in a discrete data set, the data can be grouped into classes and the tabulation and graphic representation done as for continuous data.

EXAMPLE 3.5

In an investigation of consumer behaviour and attitude regarding a particular product, consumers of frozen vegetables were asked how often they ate frozen vegetables. The frequency distribution of all the respondents' consumer behaviour is given in Table 3.13.

Table 3.13 Frequency distribution of the number of times per week that 528 respondents eat fresh vegetables (see Appendix A, Data Set 2)

Number of times per week x	Frequency (f)	Relative frequency (f/n)
0	7	0.013
1	16	0.030
2	45	0.085
3	70	0.133
4	54	0.102
5	59	0.112
6	65	0.123
7	212	0.402
	528	1.000

If the variable being studied is discrete or categorical, a visual representation of the data set can be made with the aid of a bar chart.

A **bar chart** is a graphic representation of the frequency distribution of discrete categorical data in which the values or categories are given on the horizontal axis and the frequencies are given on the vertical axis.

Figure 3.9 is a bar chart representation of the frequency distribution in Table 3.13. This figure shows unambiguously that a large percentage of consumers of frozen vegetables also eat fresh vegetables seven times per week.

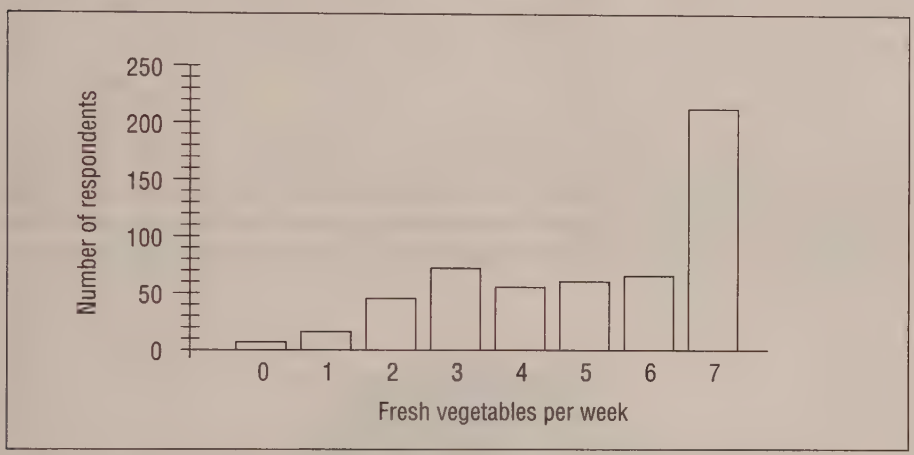


Figure 3.9 Bar chart of the number of times per week that 528 respondents eat fresh vegetables

A categorical frequency distribution indicates the number of observations in each of a number of mutually exclusive qualitative classes. Thus a share on the stock exchange can, in a given period, show a profit, remain the same or show a loss. If a portfolio of shares is studied over a certain period, the categorical frequency distribution divides these shares into the classes 'Shows a profit', 'Remains the same' and 'Shows a loss'. Table 3.14 contains a further example of a categorical frequency distribution. The distribution of percentages or ratios is also often represented graphically in the form of a pie chart.

A **pie chart** is a graphic representation of percentages by means of a circle that is subdivided into slices in such a way that the areas of these slices are proportional to the percentages.

In order to draw a pie chart, we use the fact that the total circle represents 100%. A slice representing 1% therefore comprises a central angle of $\frac{360}{100} = 3.6$ degrees. Generally $p\%$ is therefore represented by a central angle of $3.6 \times p$ degrees.

EXAMPLE 3.6

For the study on frozen vegetables the occupations of the respondents are indicated in Table 3.14. (See Appendix A, Data Set 2, V9)

Table 3.14 Frequency distribution of the type of position occupied by 529 respondents as obtained from the questionnaire on frozen vegetables

Type of position	Frequency f	Relative frequency f/n	Central angle
None	267	0.505	181.8°
Half-day	83	0.157	56.5°
Full-day	179	0.338	121.7°
	529	1.000	360°

Figure 3.10 represents a pie chart of the relative frequency distribution of Table 3.14. From the pie chart it is clear that half the respondents had no jobs, and of the respondents who did have jobs, about a third had half-day posts and two-thirds had full-day posts.

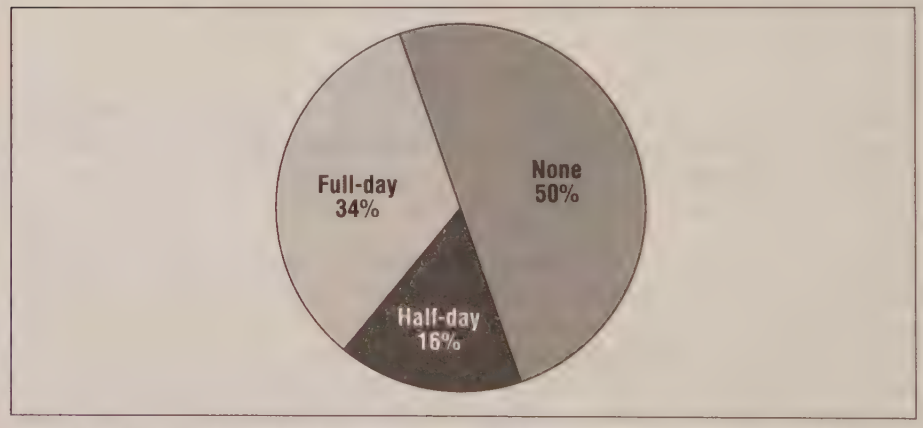


Figure 3.10 Pie chart of the type of post filled by 529 respondents as obtained from the questionnaire on frozen vegetables

9 Two-way frequency distributions and graphic representations

It is often necessary to summarize information that has a bearing on two characteristics. This summary then has to be represented visually. Such characteristics can be of a quantitative or of a categorical nature. **Two-way frequency distributions** such as those given in Tables 3.15, 3.16 and 3.18 give meaningful summaries of this type of data. The following three examples can be used to illustrate the graphic representation of two-way frequency distributions.

EXAMPLE 3.7

Table 3.15 is a two-way frequency distribution of 220 persons employed by a specific research institution, classified according to type of post and gender.

Table 3.15 Frequency distribution with respect to type of post and gender

Type of position	Men	Women	Total
Clerical	20	40	60
Research	60	80	140
Management	16	4	20
Total	96	124	220

Figure 3.11 is a horizontal bar chart representation of the data in Table 3.15. The advantage of this type of representation is that both 'percentage' and 'frequency' can be indicated on separate axes. Figure 3.11 shows clearly that the majority of posts are research posts. On a percentage basis, women occupy more clerical and research posts, while men occupy more managerial posts.

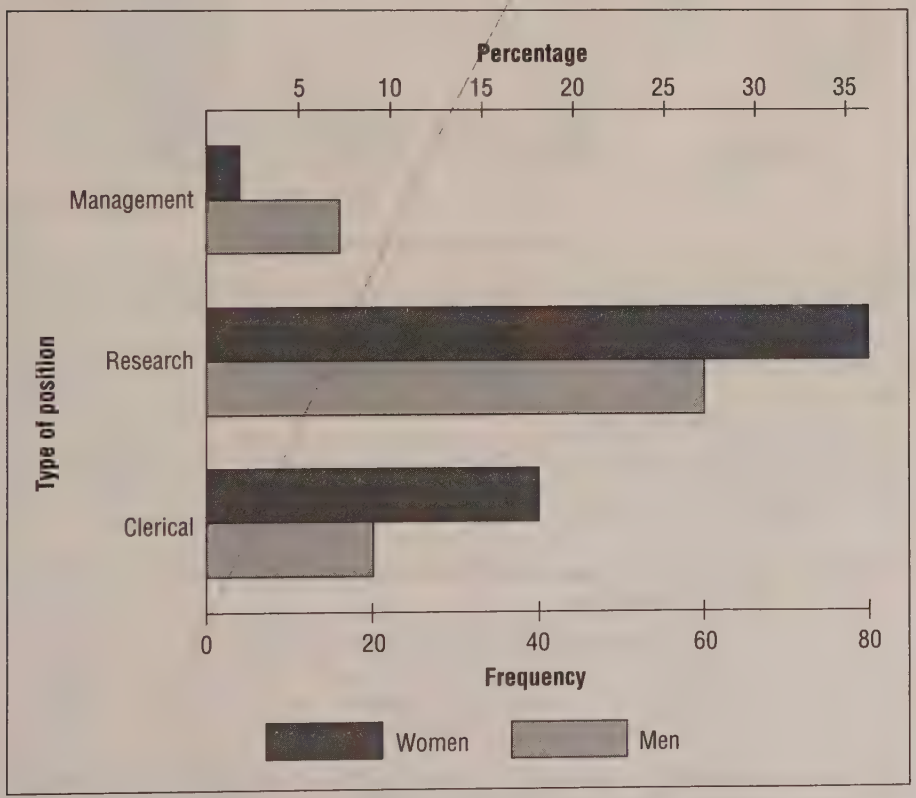


Figure 3.11 Horizontal bar chart representation with respect to type of post and gender

EXAMPLE 3.8

Table 3.16 is a two-way frequency distribution of income and the consumption of frozen vegetables. Income was divided into three categories, namely low (up to R30 000 per year), average (between R30 000 and R60 000 per year) and high (R61 000 and more per year).

Table 3.16 Frequency table of income and the consumption of frozen vegetables

Income	How often do you eat frozen vegetables?				Total
	Never	Almost never	Once or twice a month	Most weeks	
Low	14	21	29	127	191
Average	4	15	43	141	203
High	0	13	22	66	101
Total	18	49	94	334	495

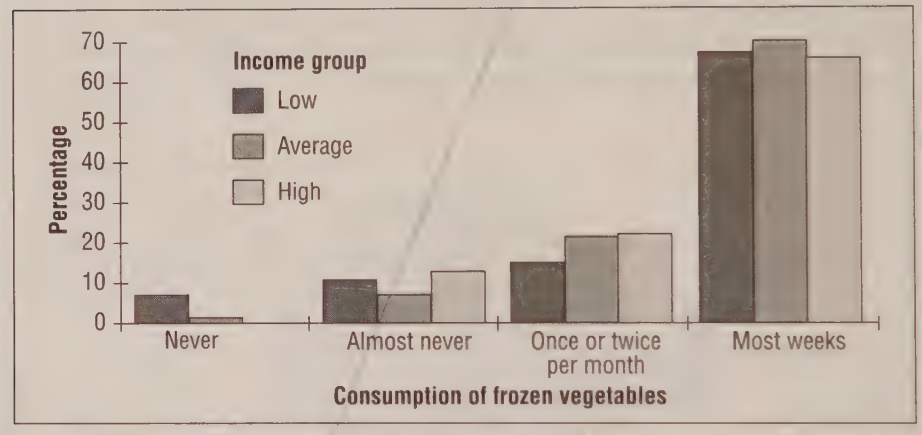


Figure 3.12 Bar chart representation of the consumption of frozen vegetables with respect to the three income groups

It is useful to express the rows in Table 3.16 as percentages of each row total. The information thus obtained enables us to determine, for example, whether the consumption pattern of respondents in the higher income group differs from that of the respondents in the other income groups. A direct comparison is not possible out of hand, since the number of respondents in each of the income categories differs. Table 3.17 is a **row percentage table** of income and consumption of frozen vegetables.

Table 3.17 Two-way row percentage table of income and the consumption of frozen vegetables

Income	How often do you eat frozen vegetables?				Total
	Never	Almost never	Once or twice a month	Most weeks	
Low	7.33	10.99	15.18	66.49	100
Average	1.97	7.39	21.18	69.46	100
High	0	12.87	21.98	65.35	100

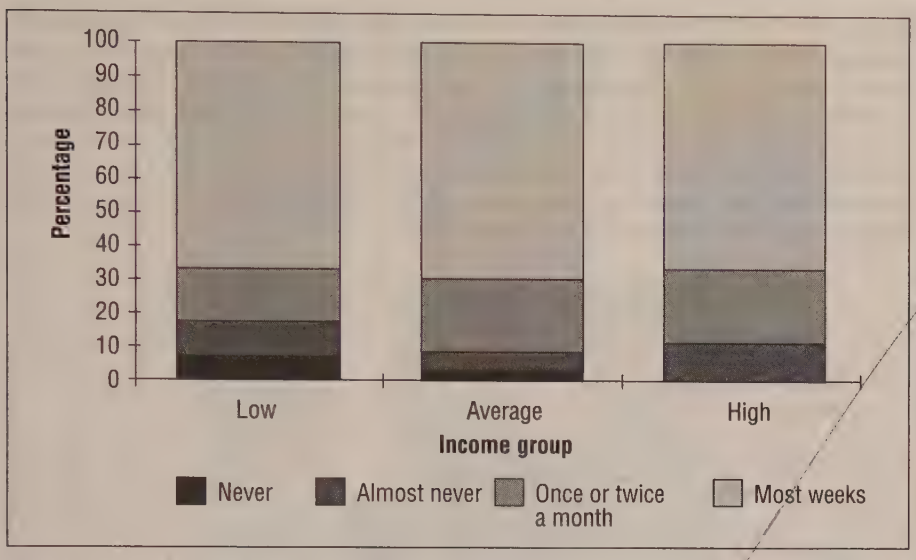


Figure 3.13 Divided bar chart representation of income and consumption of frozen vegetables

Figures 3.12 and 3.13 show two possible graphic representations of Table 3.17.

Table 3.17 and Figure 3.13 indicate that there are no meaningful differences in the consumption patterns of frozen vegetables for the three income groups. The highest percentage of consumption for all the groups is in the 'most weeks' category. Figure 3.13 is a **divided bar chart** representation of income and the percentage consumption of frozen vegetables.

EXAMPLE 3.9

An investigation was conducted to obtain information on the quality of life and working conditions of employees in the mining industry. As a result of this investigation, information was obtained on, *inter alia*, the age distribution of employees in different branches of the mining industry. Table 3.18 reflects the results of a representative sample from three branches of the mining industry, namely diamonds, coal and gold.

Table 3.18 Two-way frequency distribution of age of employees and branches of the mining industry

Age	Branch			Total
	Diamonds	Coal	Gold	
[20, 25)	12	117	477	606
[25, 35)	22	99	573	694
[35, 45)	20	95	398	513
[45, 65)	14	25	235	274
Total	68	336	1 683	2 088

Despite the fact that Table 3.18 is a condensed summary of the data set of 2 088 observations, this table contains a lot of useful information. The total column on the

right of the table gives an estimate of the age distribution of employees in the mining industry. Thus, for example, $\frac{694}{2\,088} \times 100 = 33.3\%$ of the employees are in the age group [25, 35). In contrast, the total row at the bottom of the table gives an estimate of the distribution of workers in the three main branches of the mining industry. Thus, for example, $\frac{1\,683}{2\,088} \times 100 = 80.6\%$ of the workers, i.e. the vast majority, are associated with the gold mining industry. Figure 3.14 is a **three-dimensional bar chart representation** of the distribution of the ages of the employees in three branches of the mining industry.

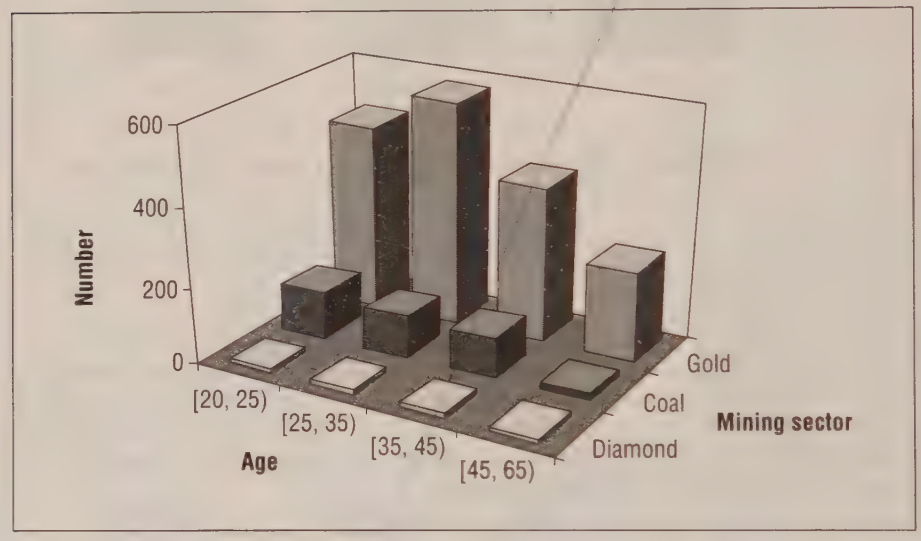


Figure 3.14 Three-dimensional bar chart of the age distribution of employees in three branches of the mining industry

From this representation it is clear that the vast majority of employees are involved in the gold mining industry, while relatively few employees in the mining industry are over the age of 45.

If Table 3.18 is expressed in terms of a column percentage table, it is possible to check whether there are any differences in age distribution across the three branches. Table 3.19 is a **two-way column percentage table** of the data in Table 3.18.

Table 3.19 Two-way column percentage distribution of age of employees and branches of the mining industry

Age	Branch		
	Diamond	Coal	Gold
[20, 25)	18.05	34.92	28.32
[25, 35)	32.20	29.46	34.06
[35, 45)	28.78	28.17	23.66
[45, 65)	20.98	7.44	13.96
	100	100	100

Figure 3.15 is a graphic representation of the data in Table 3.19. The following trend is revealed by this table and figure: on a percentage basis there are more persons under the age of 25 employed in the coal industry than in the gold or diamond industries. The opposite trend applies to persons older than 45.

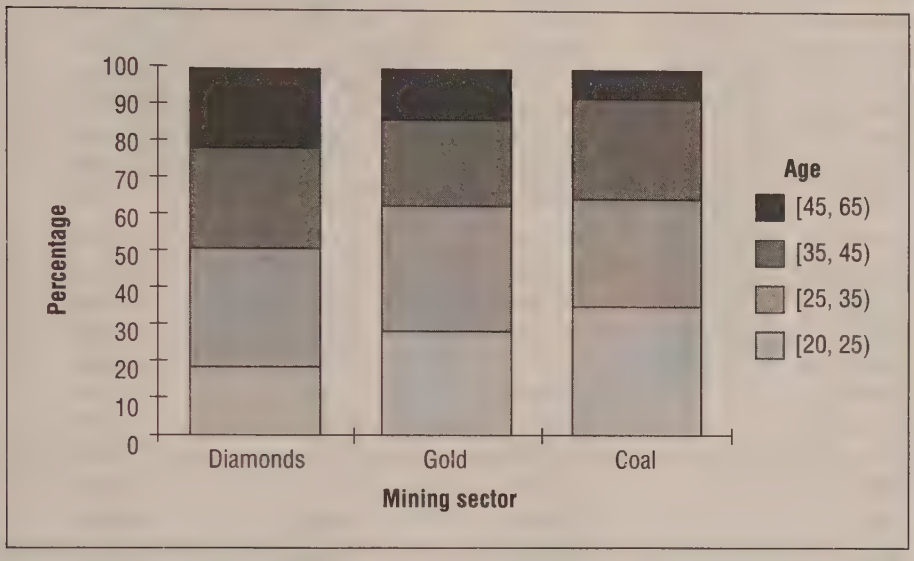


Figure 3.15 Divided bar chart representation of percentage of age distribution and branches of the mining industry

The next section contains a number of computer-generated graphic representations. These visual representations are very useful for conveying relationships and trends lucidly.

10 Using the computer for graphic representation

Because of the availability of graphics programmes, virtually any conceivable combination of standard graphical methods can be used to convey information. However, misrepresentation or representations that do not make sense, should be avoided.

EXAMPLE 3.10

In the tourist industry, information is required on the countries of origin of visitors to South Africa. The number of foreign visitors to South Africa is summarized in Table 3.20 according to country of origin. It is clear that the vast majority of foreign visitors are from Africa, with European visitors second.

Table 3.20 Foreign visitors to South Africa during 1992 according to country of origin (TBVC states excluded)

Country of origin	Number of visitors	Percentage of total	Percentage of continent
Africa	2 327 959	80.504	100.00
Botswana	240 683	8.323	10.339
Lesotho	887 811	30.702	38.137
Mozambique	88 851	3.073	3.817
Namibia	180 927	6.257	7.772
Swaziland	429 195	14.842	18.437
Zambia	26 774	0.926	1.150
Zimbabwe	392 634	13.578	16.866
Other	81 084	2.804	3.483
Europe	395 319	13.671	100.000
Germany	90 638	3.134	22.928
France	24 091	0.833	6.094
Italy	20 144	0.697	5.096
Netherlands	20 740	0.717	5.246
Portugal	5 395	0.187	1.365
Switzerland	20 300	0.702	5.135
UK	149 886	5.183	37.915
Other	64 125	2.218	16.221
Asia	67 508	2.335	100.000
Israel	11 633	0.402	17.232
Other	55 875	1.932	82.768
Americas	75 013	2.594	100.000
Canada	12 423	0.430	16.561
USA	53 871	1.863	71.816
Other	8 719	0.302	11.623
Oceania	24 821	0.858	100.000
Australia	19 582	0.677	78.893
Other	5 239	0.181	21.107
Unspecified	1 101	0.038	
Total	2 891 721	100.000	

Source: *Bulletin of Statistics*, Central Statistical Service, June 1993

The information contained in Table 3.20 can be represented graphically by a pie chart as in Figure 3.16.

EXAMPLE 3.11

Table 3.21 reflects an estimated distribution of the larger mammals of the Kruger National Park. Advanced statistical techniques are used to determine animal populations in the Kruger Park, since it is impossible to determine the numbers exactly.

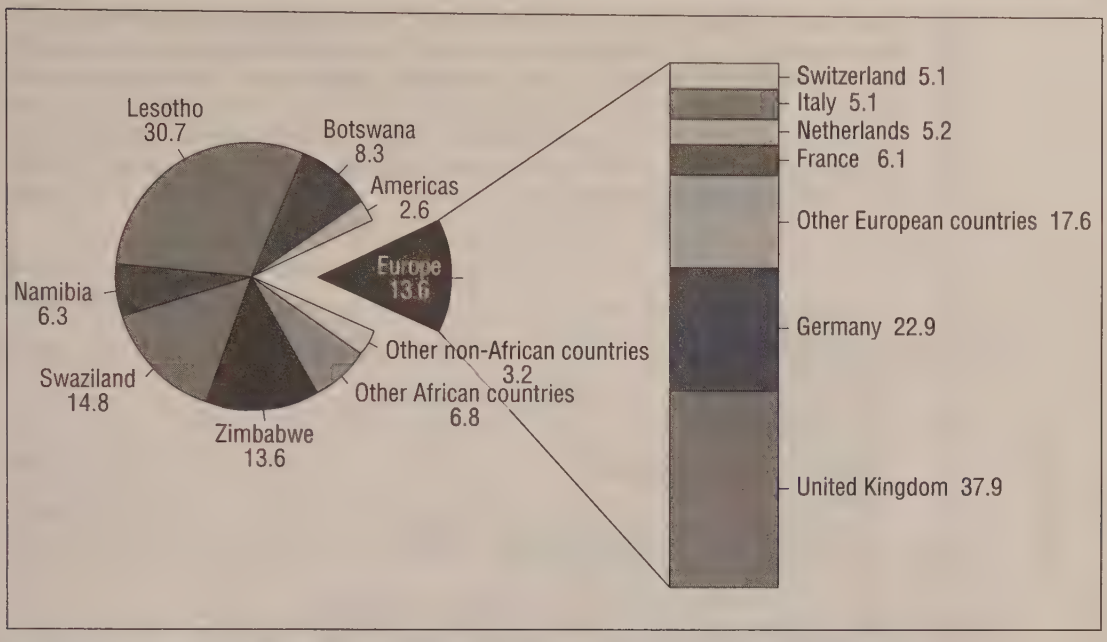


Figure 3.16 Pie chart of foreign visitors to South Africa during 1992 (TBVC states excluded)

Table 3.21 Estimated numbers of the most important animal species in the Kruger National Park

Species	Number	Percentage of total
Common animals	212 600	89.155
Impala	128 000	53.678
Quagga	33 000	13.839
Buffalo	27 500	11.532
Blue wildebeest	13 600	5.703
Kudu	10 500	4.403
Less common animals	21 600	9.058
Elephant	7 500	3.145
Giraffe	5 500	2.306
Waterbuck	4 000	1.677
Hippo	2 500	1.048
Sable	2 100	0.881
Rare animals	4 260	1.786
Lion	1 500	0.629
Square-lipped rhinoceros	1 160	0.486
Leopard	750	0.315
Roan	425	0.178
Cheetah	275	0.115
Hook-lipped rhinoceros	150	0.063
Total	238 460	100.000

A problem surfaces when the data in Table 3.21 is represented graphically. The extensive numbers of common animals dwarf the distribution of the rare animals. The data should therefore preferably be sorted into groups of, say, most common animals, less common animals and rare animals, as shown in Table 3.21. Figure 3.17 shows a three-dimensional bar chart of the rare large animal species in the Kruger National Park. In this figure animal pictures could be used to enhance the effect of the graphic representation.

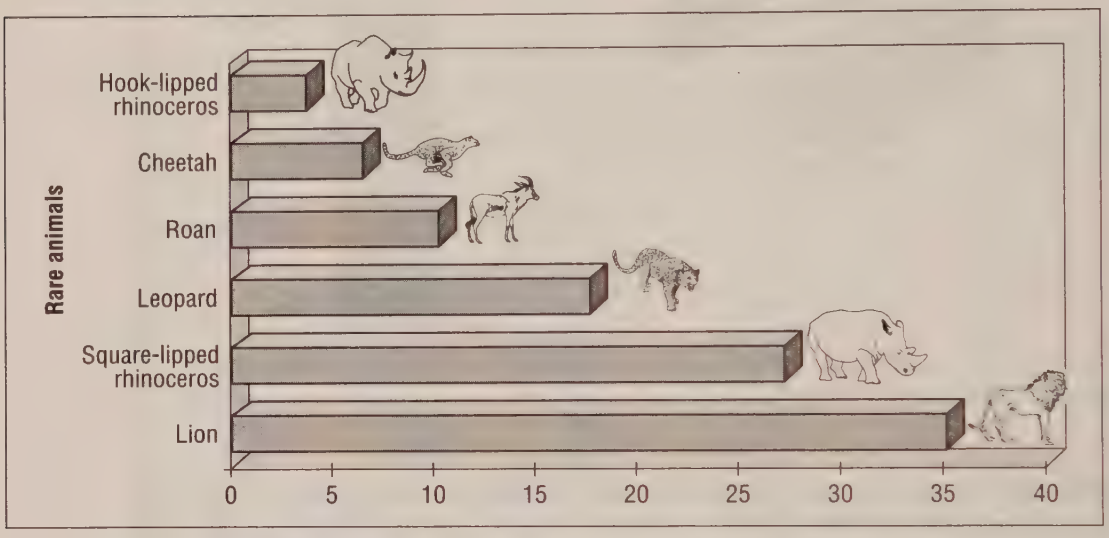


Figure 3.17 Bar chart of rare large animal species in the Kruger National Park

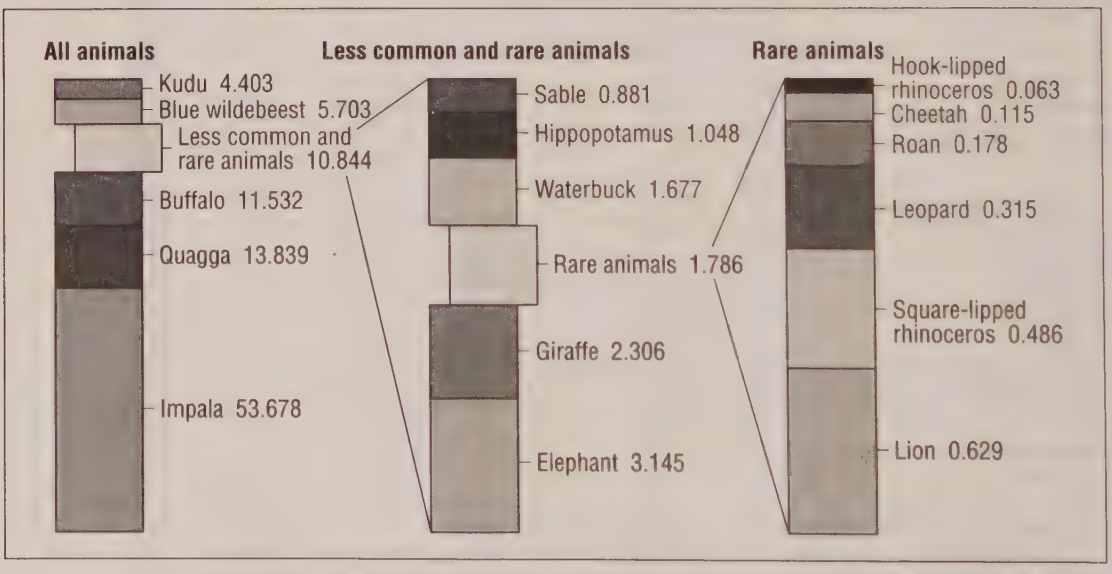


Figure 3.18 Combined bar chart (percentage) of large animal species in the Kruger National Park

Figure 3.18 shows a combined bar chart of common, less common and rare large animal species in the game park. A further division into subgroups clearly reflects the division within each subgroup, with the advantage that the detail is not lost in the larger picture.

EXAMPLE 3.12

The 1991 census reveals interesting demographic characteristics for the respective population groups, *inter alia*, that urbanization is distributed as demonstrated in Table 3.22.

Table 3.22 Urbanization and population composition in South Africa in 1991

	Whites	Coloureds	Asians	Blacks	Total
Urban %	91.2	83.3	96.2	47.1	60.3
Non-urban %	8.8	16.7	3.8	52.9	39.7
Population %	16.3	10.4	3.1	70.2	

Source: *Demographic Statistics 1992*: Central Statistical Service
Total population: 30 797 000

The three-dimensional pie chart in Figure 3.19 shows a graphic representation of the population composition. When using three-dimensional pie charts, remember that the front segment can confuse the reader because the three-dimensional effect artificially enlarges the surface area.

Figure 3.20 shows a graphic representation of the urbanization of the population by population group.

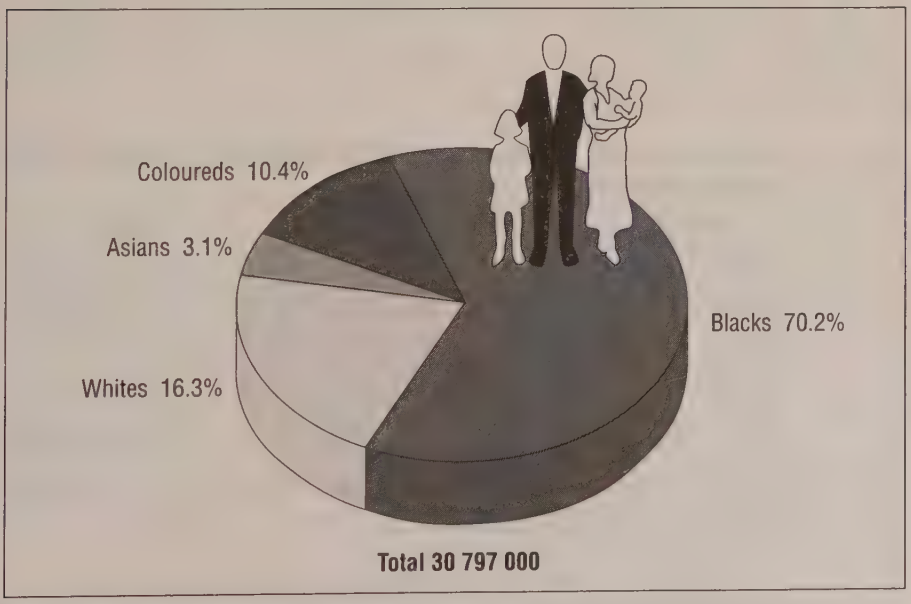


Figure 3.19 Three-dimensional pie chart representation of the population composition
Source: *Demographic Statistics 1992*: Central Statistical Service

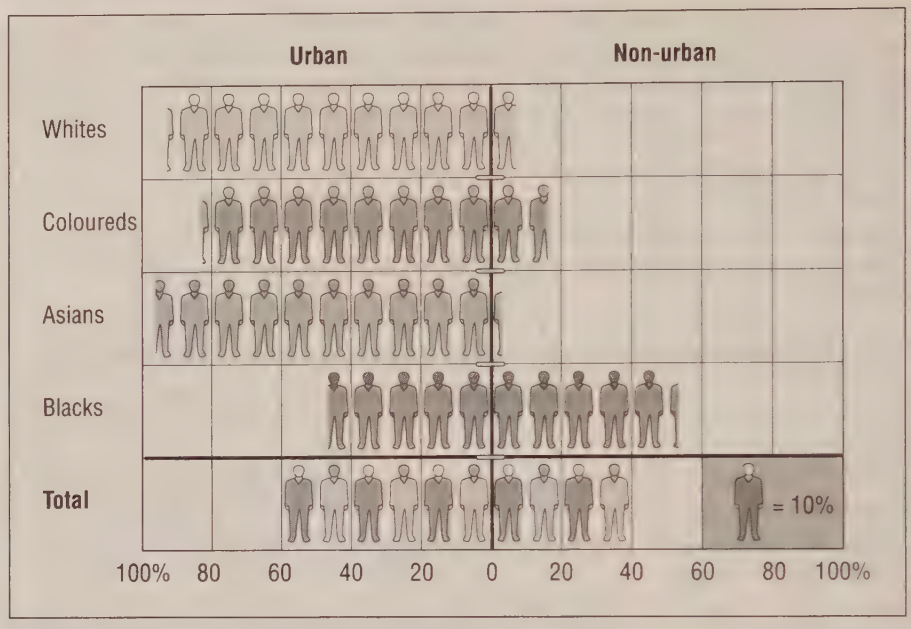


Figure 3.20 Urbanization of the population groups in South Africa as in 1991
Source: *Demographic Statistics 1992*, Central Statistical Service

EXAMPLE 3.13

During a census a wide range of demographic information is collected, *inter alia*, the age of each individual in South Africa. Determination of the population distribution is essential for purposes of planning social services such as education, health services and care of the aged. Table 3.23 shows the age distribution according to gender and population group.

Table 3.23 Age distribution of RSA population according to gender and population group

	Whites		Coloureds		Asians		Blacks	
Age	Male	Female	Male	Female	Male	Female	Male	Female
0 – 4	164 808	158 585	153 765	153 684	38 431	36 946	989 632	1 000 809
5 – 19	582 507	558 176	488 196	483 970	139 957	136 354	3 001 952	3 108 785
20 – 34	543 845	553 104	394 286	428 496	111 624	117 239	2 464 580	2 577 414
35 – 49	465 166	468 307	224 047	250 424	83 733	88 807	1 404 206	1 361 228
50 – 64	294 007	310 288	114 386	133 590	39 535	44 043	625 157	750 555
65 +	170 679	252 401	42 716	61 762	12 327	14 880	284 474	404 528
Total	2 221 012	2 300 861	1 417 396	1 511 926	425 607	438 269	8 770 001	9 203 319

Source: *Demographic Statistics 1992*, Central Statistical Service

A graphic representation can be drawn with the aid of age pyramids as in Figure 3.21.

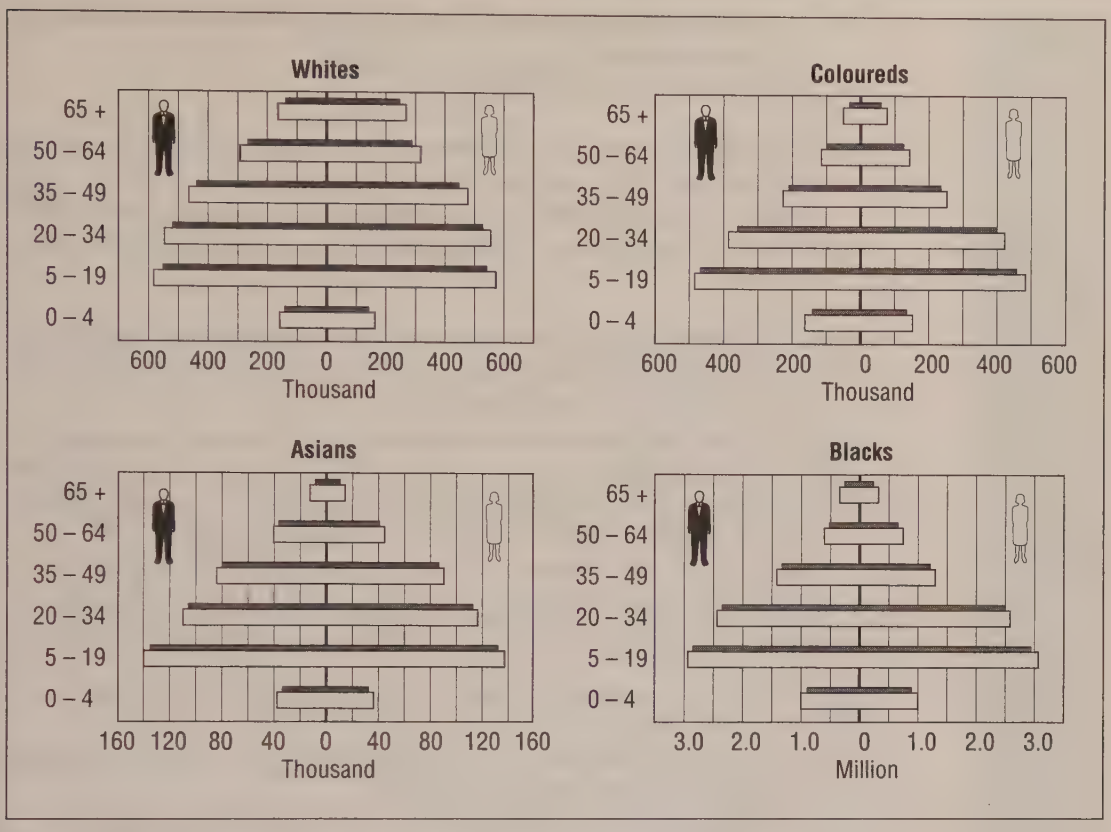


Figure 3.21 Age pyramids for the respective population groups according to gender

11 Concluding remarks

Data analysis is often preceded by descriptive analysis of the underlying distribution of the relevant variables. In such cases, graphic representations can be useful for estimating at a glance the nature of the distributions and relationships. In Chapters 4 and 5 we will show how calculations can also be used to summarize data sets.

Exercises

1 (a) Use the floor area in Column 5 of Appendix A, Data set 1 to compile a data array for the floor areas of houses 1 to 60 in Durban.

(b) Determine the minimum and maximum values. Assume that you are required to construct a frequency distribution for the data in (a). Use Sturge's rule to determine a possible number of class intervals and their lengths.

(c) Construct a frequency distribution for the floor area of the first 60 houses in Durban.
(d) Draw a histogram of the floor area.

2 (a) Construct a frequency distribution for the selling prices of the first 100 houses in Pretoria (see Appendix A, Data set 1, column 3). Use a class length of R50 000 and start the first class boundary at R50 000.
(b) Draw a histogram and a frequency polygon for the frequency distribution.

3 (a) Construct a frequency distribution for the selling prices of houses 241 to 300 in Johannesburg. Use a class length of R50 000 and start the first class boundary at R50 000.
(b) Represent this information graphically in the form of a histogram.
(c) Draw a cumulative frequency polygon. Remember that you must first determine the cumulative frequency distribution. Use your graphic representation to find answers for the following questions:
(i) How many houses have a selling price of less than R100 000?
(ii) How many houses have a selling price of at least R175 000?
(iii) The selling price of half of the houses is y. Determine the value of y.

4 (a) Construct a combined relative frequency distribution for the data in Exercises 2 and 3.
(b) Draw, on the same axes, the relative frequency polygons based on the relative frequency distribution of (a).
(c) What conclusions can you draw from this graphic representation?
(d) Construct a combined relative frequency distribution based on (a).
(e) Draw cumulative relative frequency polygons based on (d).
(f) What conclusions can you draw from this graphic representation?
(g) Which graphic representation is preferable, (b) or (e)?

5 (a) Construct a categorized frequency distribution for the plot size given in Appendix A, Data Set 1, column 6, for the first 50 houses in each of the four cities.
(b) Prepare a graphic representation using stacked bar charts.
(c) Write a short summary of your conclusions drawn from the graphic representations.

6 (a) Construct a frequency distribution for the ages of the first 100 houses in each of the cities (see Appendix A, Data Set 1, column 7).
(b) Draw suitable bar charts for these distributions.
(c) Write a short summary of your conclusions drawn from the graphic representations.

7 (a) Construct a relative frequency distribution for the plot sizes of houses 1 to 250 in Johannesburg as in Appendix A, Data Set 1, column 6.
(b) Draw a pie chart of the percentage of houses with different-sized plots.

8 The following table shows the weekly salaries of 72 clerks.

Class	Weekly salary (rand)	Frequency
1	[200, 300)	12
2	[300, 400)	25
3	[400, 500)	17
4	[500, 750)	10
5	[750, 1 000)	6
6	[1 000, 1 500)	2

Draw a histogram of the data. Note that the length of the classes differs. Since the areas of the rectangles of the histogram must be proportional to the frequencies, the heights of the rectangle that coincides with the fourth class is taken as $\frac{100}{250} \times 10 = 4$ units and the one that coincides with the sixth class as $\frac{100}{500} \times 2 = 0.04$. The vertical axis is omitted, and the frequency that coincides with each of the classes is indicated above the rectangle at the corresponding class.

9 The heart rate of thirty randomly selected physical education students is measured after ten seconds of light exercise and noted as follows (heart beat per minute).

82 95 92 62 85 92 82 95 70 85 84 95 91 82 94
76 88 91 87 80 68 58 76 85 71 60 75 88 64 74

(a) Prepare an array of this data.
(b) Apply Sturge's rule to determine a meaningful number of class intervals.
(c) Calculate the class width.
(d) Construct a frequency distribution for the abovementioned data.
(e) Draw the histogram for the frequency distribution referred to in (d).
(f) Draw the frequency polygon on the graph referred to in (e).

10 The following is a frequency distribution of the annual loss (R1 000) sustained by forty chain stores as a result of shoplifting:

Loss (R1 000)	Number of shops
[4.5, 9.5)	3
[9.5, 14.5)	14
[14.5, 19.5)	11
[19.5, 24.5)	10
[24.5, 29.5)	2

(a) Draw the histogram and frequency polygon from the frequency distribution.
(b) Construct the frequency distribution and represent it graphically.
(c) How many shops showed a loss of less than R14 000?
(d) Construct the relative frequency distribution.

11 An investigation is being conducted on the seriousness of injuries sustained in 100 successive accidents which occurred in an industry.

Minor	60
Light	32
Severe	8

Draw a bar chart for this information.

12 The table below gives the frequency distribution for the heights (cm) of 120 women.

Heights	Number of women
[138, 144)	2
[144, 150)	2
[150, 156)	10
[156, 162)	26
[162, 168)	28
[168, 174)	27
[174, 180)	16
[180, 186)	6
[186, 192)	3

(a) Draw a histogram for the frequency distribution.
(b) Draw the frequency polygram on the graph referred to in (a).
(c) Construct the cumulative frequency distribution and represent it graphically.
(d) Calculate the relative frequencies.
(e) Draw the relative frequency histogram.

13 The smallest observation in a data set comprising 500 observations (continuous data) is 30.45, and the largest 356.72. Determine a suitable set of class intervals for the construction of a frequency distribution. (Apply Sturge's rule to determine the number of classes.)

14 If the data set in Exercise 13 were to consist of 15 000 observations, determine a suitable set of class intervals.

15 Consider the following frequency distribution for points scored by the winning team in 100 rugby test matches:

Scores	Frequency
[3, 9)	18
[9, 15)	30
[15, 21)	21
[21, 27)	12
[27, 33)	10
[33, 39)	7
[39, 45)	2

(a) Calculate the class marks.
(b) Draw a histogram and a frequency polygon.
(c) Construct the cumulative frequency distribution.
(d) Draw the cumulative frequency polygon.
(e) In how many of the test matches did the winning team score less than eighteen points?

16 The monthly expenditure on electricity (rand/month) of 66 families is drawn for purposes of an investigation into electricity supply.

315	276	375	236	217	226	445
216	238	259	296	245	233	386
286	257	327	183	324	366	357
216	477	365	234	245	463	326
347	319	278	259	276	496	246
376	316	186	296	216	426	195
466	178	236	239	536	237	325
234	163	256	268	216	286	386
355	375	366	287	297	266	275
436	445	447				

(a) Construct a frequency distribution for this data.
(b) Construct the cumulative frequency distribution.
(c) Draw the histogram and the frequency polygon.
(d) Draw the cumulative frequency polygon and answer the following questions:
  (i) How many families spend less than R275 per month on electricity?
  (ii) How many families spend more than R300 per month on electricity?
  (iii) How many families spend between R200 and R250 per month on electricity?

17 The following table represents the average weekly amount (rand) received from the private sector by registered South African charity organizations.

Amount	Number of organisations
[250, 500)	2
[500, 750)	3
[750, 1 000)	3
[1 000, 1 250)	7
[1 250, 1 500)	12
[1 500, 2 000)	19
[2 000, 3 000)	9
[3 000, 4 000)	2
[4 000, 5 000)	1
[5 000, 10 000)	8

(a) Present this information in the form of a histogram.
(b) Draw the cumulative frequency polygon.
(c) How many of these organizations received less than R2 500 a week?

18 Assume that the following information is given with regard to 250 houses in a specific suburb:

Number of bedrooms	Outside walls plastered	Outside walls unplastered
2	18	2
3	108	72
4 and more	10	40

(a) Draw a bar chart for the frequency of the number of bedrooms.
(b) Present the information given in columns 2 and 3 in graphic form using
  (i) a pie chart;
  (ii) a bar chart.
(c) Present the information from the above table in graphic form in accordance with the method illustrated in Example 3.8.

19 In a survey conducted at a company, employees were divided into the following age groups.

Age	Frequency
15 – 19	0
19 – 24	16
25 – 34	37
35 – 44	54
45 – 54	53
55 – 64	11
65 +	0

(a) Present this information in graphic form by means of a histogram.
(b) Draw the frequency polygon over the histogram in (a).
(c) What do you notice with regard to the ages of employees?

20 The years of service of employees in a business is divided as follows:

Years of service	Frequency
Less than 1	4
1 – 3	25
4 – 10	36
11 – 20	61
21 – 30	33
30 +	12

(a) Draw a histogram of the above data.
(b) Calculate the cumulative relative frequencies for the years of service.
(c) Draw a cumulative relative frequency polygon for the data.
(d) Half the employees have more than how many years of service?
(e) What percentage of employees have been with the company for longer than fifteen years?

21 The monthly income of employees in a company is divided as follows:

Monthly earnings	Frequency
R1 000 – R1 499	23
R1 500 – R1 999	33
R2 000 – R2 499	22
R2 500 – R2 999	11
R3 000 – R3 499	13
R3 500 – R3 999	8
R4 000 – R4 499	15
R4 500 – R4 999	4
R5 000 +	31

(a) Calculate cumulative relative frequencies for the data.
(b) Draw a cumulative relative frequency polygon for the data.
(c) What percentage of employees earn R4 000 and more per month?

22 Staff numbers at the Central Statistical Service are compared with staff numbers at similar statistics offices in other countries. In Table 3.20 it is important to take into account the population figures of the respective countries.

Staff numbers in statistics offices: comparison between South Africa and other countries

	Staff numbers	Population (million)	Staff per million
South Africa (1992)	614	32	19
Australia	3 500	16	219
France	7 000	54	130
Hungary	8 000	11	727
Canada	4 200	22	191
Netherlands	2 600	14	186
Britain (1990)*	4 200	57	74
America (1979)*	30 000	225	133

*Countries with decentralized systems. Totals for all the branches are given.

Source: *Finansies & Tegniek*, 5 June 1992.

(a) Use suitable graphic representations to reflect the population sizes of the respective countries.
(b) Use suitable graphic representations to reflect the staff numbers per million of the population.

23 A specific component is used in manufacturing robot arms. The number of components per order for a company has been collected over three years. The ordered data is given in the following table.

Components per order			Components per order		
1989	1990	1991	1989	1990	1991
191	156	161	304	380	694
200	162	161	314	411	701
200	162	171	317	421	704
200	165	171	317	484	717
203	173	177	321	500	755
213	178	179	326	507	761
214	180	179	327	522	780
219	181	181	331	524	789
219	183	181	335	555	822
219	184	181	339	557	826
224	192	185	343	560	829
226	198	198	345	593	838
231	201	198	375	626	844
234	203	210	392	628	917
234	206	218	428	687	979
236	210	220	475	727	983
238	217	223	517	743	1 003
241	218	223	528	763	1 041
246	219	224	543	818	1 093
251	220	225	607	839	1 112
258	228	228	632	865	1 168
264	228	240	635	869	
267	235	244	652	895	
273	239	245	656	919	
273	242	247	658	990	
274	245	260	662	98	
274	249	261	709	1 028	
275	249	280	757	1 129	
280	257	365	802	1 193	
282	258	459	846	1 233	
283	262	522	874	1 253	
285	263	547	887	1 326	
288	268	553	896		
289	271	565	902		
291	272	617	905		
291	290	626	923		
293	290	627	1 004		
296	290	651	1 005		
301	292	653	1 119		
303	313	687	1 135		

Data obtained from MBL group JHB01192

(a) Construct suitable frequency distributions to compare the distribution of components per order over the three years.

(b) Construct suitable graphic representations of the data.

(c) How do the sales volumes compare for the three years?

24 The amount of beer produced by a brewery is given in Appendix A, Data Set 5.
   (a) Construct frequency distributions to compare the amount of beer produced for the two financial years 1992 and 1993.
   (b) Calculate the cumulative frequency distributions.
   (c) Construct the histograms and frequency polygons.
   (d) Construct the cumulative frequency polygons and answer the following questions for each financial year:
      (i) For how many weeks was the amount of beer produced greater than 15 000 hℓ?
      (ii) For how many weeks was the amount of beer produced greater than 20 000 hℓ?
      (iii) What amount did the produced amount exceed for 40 out of the 53 weeks?
   (e) What conclusions can you draw concerning the amount of beer produced in the two financial years under investigation?

25 The amount of electricity used by a brewery is reflected in Appendix A, Data Set 5. Use this variable in answering the following questions:
   (a) Construct relative frequency distributions to compare the amount of electricity used for the two financial years 1992 and 1993.
   (b) Calculate the cumulative relative frequency distributions.
   (c) Construct the relative frequency histograms and frequency polygons.
   (d) Construct the cumulative relative frequency polygons and answer the following questions for each financial year:
      (i) What percentage of the time was the amount of electricity used less than 200 000 kWh?
      (ii) What percentage of the time was the amount of electricity used greater than 250 000 kWh?
      (iii) What electricity usage level was exceeded half of the time?
      (iv) What electricity usage level was exceeded three quarters of the time?
   (e) How does the amount of electricity used compare for the 1992 and 1993 financial years?

26 The amount of electricity used (kWh/hℓ) to produce each hectolitre of beer is reflected in Appendix A, Data Set 5. Use this variable in answering the following questions:
   (a) Construct frequency tables to compare the two financial years.
   (b) Calculate the cumulative frequency distributions.
   (c) Construct the cumulative relative frequency distributions.
   (d) What conclusion can be reached concerning the amount of electricity used to produce a hectolitre of beer for the two financial years?

27 The city engineer of a certain municipality is responsible for approval of building plans. Since time is a critical factor in development projects, the method followed for approving these building plans must be effective. An

investigation of the time taken to approve building plans is decided on. A sample of 100 approved building plans is drawn and the number of working days taken for approval is noted. After initialization of changes in the approval process, another sample of 100 building plans is drawn to test whether the new method of approval is more effective. The number of working days which pass before the plans are approved for both the old and the new methods are given in ordered form below.

Number of days	Frequency for old procedure	Frequency for new procedure
1	2	1
2	2	8
3	0	18
4	1	8
5	0	11
6	0	4
7	2	11
8	2	6
9	1	5
10	5	3
11	6	3
12	2	4
13	1	2
14	2	4
15	5	3
16	8	0
17	2	3
18	3	1
19	7	0
20	4	0
21	2	5
22	2	0
23	1	0
24	2	0
25	2	0
26	2	0
28	6	0
29	1	0
30	6	0
31	4	0
32	2	0
33	1	0
34	2	0
35	2	0
38	4	0
40	1	0
42	2	0
43	2	0
47	1	0
Total	100	100

Data obtained from MBL group CTN00193

(a) Use the following classes to construct frequency distributions for the time taken before building plans are approved for both the old and the new method:
1 – 5
6 – 10
11 – 15
16 – 20
21 – 25
26 – 30
31 – 35
36 – 40
41 – 45
46 – 50

(b) Construct frequency polygons for the distributions in (a).

(c) Construct cumulative frequency distributions for the time taken before building plans are approved for both the old and the new method.

(d) Construct cumulative frequency polygons for both the old and the new method.

(e) What conclusions can be drawn from the graphic representations?

28 The bed occupancy of a 350 bed hospital must be compared for February and March. The following table reflects the daily bed occupancy for both months.

February 1992				March 1992			
1	261	17	275	1	302	17	294
2	271	18	293	2	304	18	300
3	296	19	277	3	320	19	296
4	308	20	274	4	297	20	303
5	325	21	282	5	285	21	300
6	294	22	282	6	285	22	301
7	287	23	297	7	269	23	329
8	270	24	307	8	266	24	334
9	273	25	301	9	287	25	335
10	297	26	318	10	292	26	328
11	308	27	289	11	308	27	300
12	308	28	279	12	306	28	281
13	283	29	288	13	300	29	284
14	284			14	288	30	285
15	262			15	274	31	301
16	263			16	304		

(a) Construct frequency distributions for the bed occupancy for each month.

(b) Calculate the cumulative frequency distribution for each month.

(c) Construct a graphic representation for the distributions in (b).

(d) How does the bed occupancy compare for the two months under investigation?

29 Use the respondents numbered 101 to 175 in the frozen vegetable investigation, Appendix A, Data Set 2, and construct frequency distributions for

the number of cooked lunches (V10a) and dinners (V10b). Use suitable techniques to represent these observations graphically.

30 Construct a frequency distribution and a graphic representation for the total yearly incomes (V11) for respondents 1 to 100 in the frozen vegetable investigation as reflected in Appendix A, Data Set 2.

31 Construct a frequency distribution and graphic representation for the variable Know for respondents 1 to 100 in the frozen vegetable investigation as reflected in Appendix A, Data Set 2.

32 Use the data pertaining to Health1 and Health2 for respondents 301 to 350 from Appendix A, Data Set 2.
(a) Construct relative frequency distributions.
(b) Construct relative frequency histograms.
(c) Construct cumulative relative frequency distributions.
(d) Construct the cumulative relative frequency polygons.
(e) What conclusions can be derived from these analyses?

33 Repeat Exercise 32 for the variables Conv1 and Conv2 for respondents 301 to 350.

34 Repeat Exercise 32 for the variables Finan1 and Finan2 for respondents 301 to 350.

35 Repeat Exercise 32 for the variables Diff1, Diff2 and Diff3 for respondents 301 to 350.

36 The following frequency distributions give the monthly earnings of a sample of fulltime workers by gender. Compare the two distributions and comment on the differences. Use appropriate graphs.

Earnings	Class mark	Women	Men
[0, R500)	250	98	129
[R500, R1 000)	750	127	203
[R1 000, R1 500)	1 250	326	476
[R1 500, R2 000)	1 750	185	354
[R2 000, R3 000)	2 500	81	143
[R3 000, R4 000)	3 500	27	72
[R4 000, R5 000)	4 500	10	31
[R5 000, R8 000)	6 500	3	21

37 A new chicken farm that uses modern technology is compared to an established chicken farm. The information reflected in the following table gives the production per hen (number of eggs laid per hen per week) and the deaths per week (the number of deaths per 1 000 hens per week) for the two farms for a period of twenty weeks.

Week	Established poultry farm		New poultry farm	
	Production	Deaths	Production	Deaths
1	6.02	0.83	5.74	1.40
2	6.06	0.90	5.67	0.98
3	6.02	0.86	5.63	1.38
4	5.95	1.27	5.80	1.56
5	5.88	1.55	5.83	1.74
6	5.31	1.16	5.11	1.85
7	5.47	1.46	5.33	1.54
8	5.78	1.39	5.58	1.24
9	4.84	1.52	5.09	1.32
10	6.03	1.45	5.55	1.09
11	6.57	0.70	5.39	0.86
12	6.53	0.77	5.54	0.90
13	6.57	0.80	5.49	0.82
14	6.53	0.92	5.46	0.93
15	6.48	0.89	5.44	0.97
16	5.79	1.98	6.30	0.67
17	5.80	1.39	6.33	0.71
18	5.75	0.99	6.28	0.73
19	5.73	1.38	6.27	0.75
20	5.65	1.37	6.26	0.77

Data obtained from MBL group JHB00992

(a) Construct frequency distributions to compare the production of the two farms.
(b) Construct histograms for the two farms' production.
(c) How do the two farms compare with regard to production?

38 Repeat Exercise 37 for the deaths per 1 000 hens.

CHAPTER 4

Descriptive measures of location or central tendency

1 Introduction

Compiling a frequency table for an observed data set and the graphic representation of the distribution thus obtained can be regarded as a first step in analysing the data. This table provides an overall picture of the data set of, for example, information on the order of magnitude and the range of the observations. However, such an overall picture gives only an approximate indication of specific properties such as the midpoint and spread of the data. When analysing a data set there are at most four different properties of shape that interest the investigator, namely **location**, **spread**, **symmetry**, and **kurtosis** or **peakedness**.

Although neither the histogram nor the frequency polygon of a data set is a smooth curve, the general or basic shape of either could be represented by a smooth curve, as indicated, for example, in Figure 4.1. This curve indicates the shape or distribution of the data set.

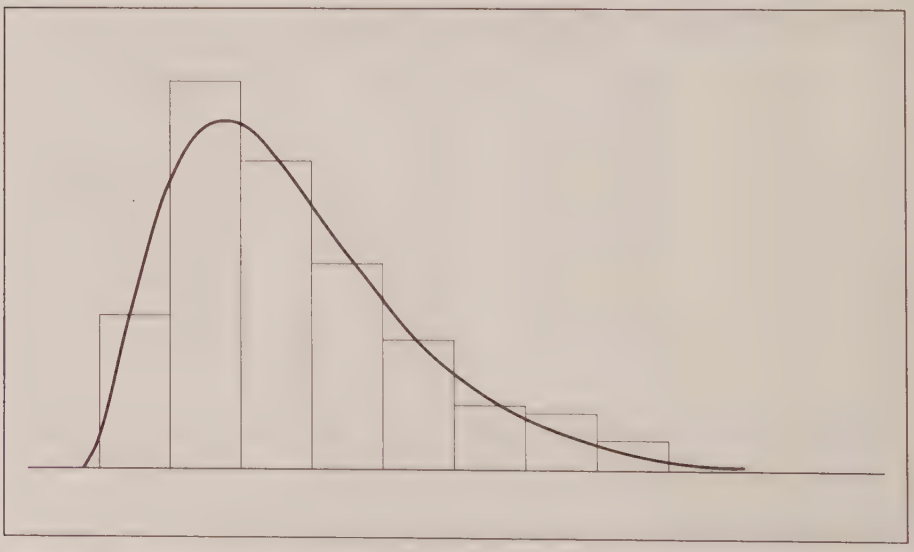


Figure 4.1 Approximation of a histogram by a smooth curve

Each of these properties of shape can be described by a single numerical value. These numerical values are known as descriptive measures. There are descriptive measures of location, spread, symmetry and kurtosis. Before analysing a data set, it is important to know what kind of data one is dealing with – continuous, discrete, nominal or ordinal (see Chapter 1, Section 3). Descriptive measures can be calculated and interpreted meaningfully only if the unique nature and the measuring unit of the data concerned are taken into account. Each of these kinds of measures will now be discussed.

1.1 Descriptive measures of location

A measure of location or position gives an indication of the 'midpoint' or general size of the distribution. In Figure 4.2 distributions *A* and *B* correspond with regard to location, whereas distribution *C*'s location is larger than those of both *A* and *B*.

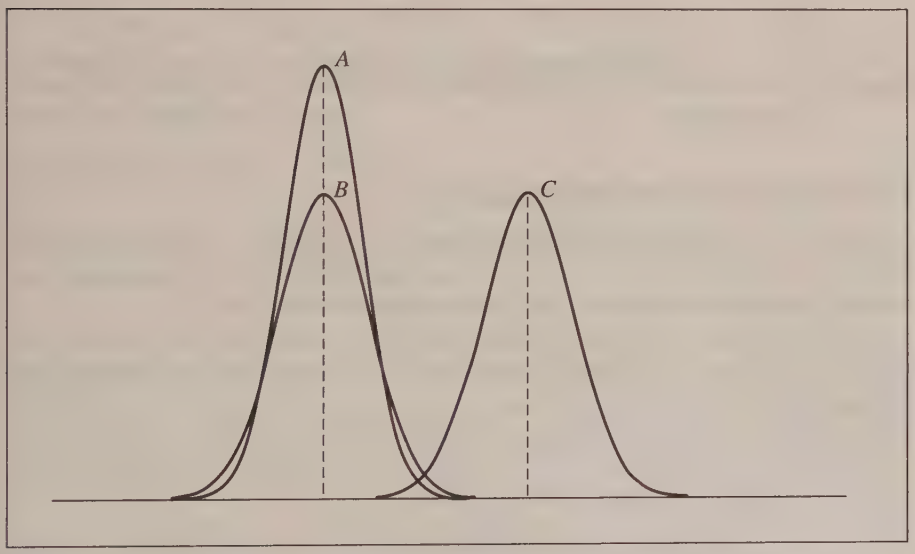


Figure 4.2 Comparison of properties of location and spread of three distributions

1.2 Descriptive measures of spread

A measure of spread measures the inherent variation of the observations in a data set. In Figure 4.2 distributions *B* and *C* correspond to some extent with regard to spread whereas distribution *A* shows a smaller spread.

1.3 Descriptive measures of symmetry

A distribution is symmetrical if a vertical line divides it into two equal parts so that each part is the mirror image of the other. All three distributions in Figure

4.2 are symmetrical, while in Figure 4.3, distribution A is negatively skewed and distribution B is positively skewed. A symmetry measure indicates the degree of skewness of a distribution.

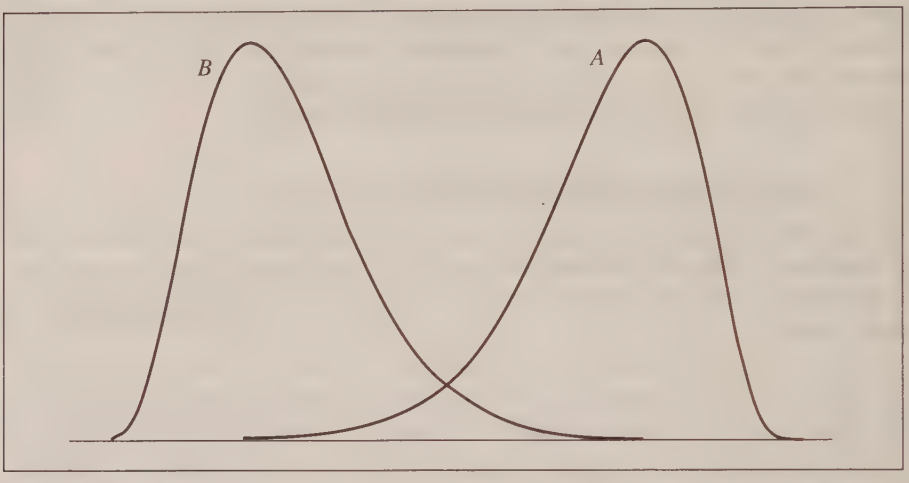


Figure 4.3 Representation of a negatively skewed distribution A and a positively skewed distribution B

1.4 Descriptive measures of kurtosis

A measure of kurtosis indicates the measure of peakedness of a distribution. In Figure 4.4 distributions A and B correspond more or less with regard to location, spread and symmetry, but A has a bigger kurtosis than B.

Chapters 4 and 5 will concentrate mainly on measures of location and spread. At the end of Chapter 5 measures of symmetry and kurtosis will be discussed briefly.

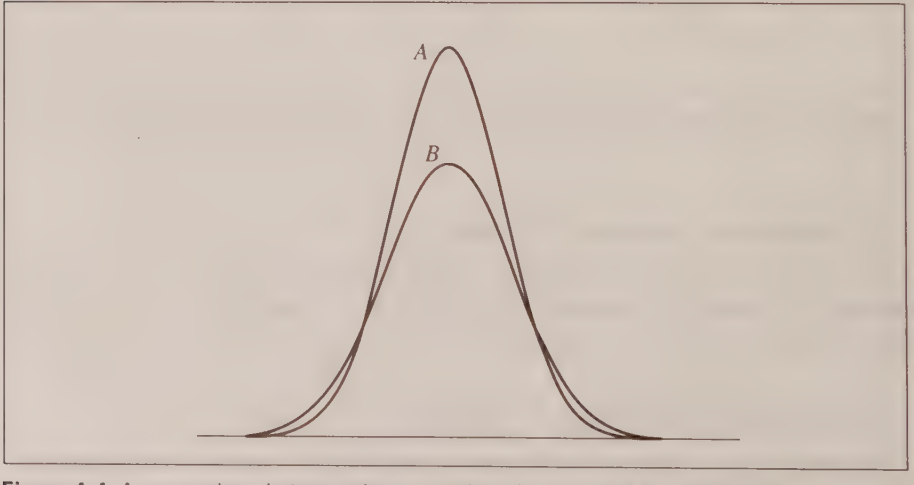


Figure 4.4 A comparison between the properties of kurtosis of two distributions

Notation

The mathematical notation used in this textbook will now be discussed briefly. Individual observations of a particular variable, for example, x, y or z, are indicated by subscript letters or figures. Thus $x_1, x_2, \ldots, x_n$ indicate n different observations of the variable x; $y_1, y_2, \ldots, y_k$ indicate k different observations of the variable y, and so forth. In a raw data set the subscripts usually refer to the sequence in which the data has been collected. In this chapter the variable that is being considered will consistently be indicated by x and the sample of observations by $x_1, x_2, \ldots, x_n$ or simply by x_i, where $i = 1, 2, \ldots, n$. The sum of the observations is indicated by $\sum x_i$ or simply by $\sum x$, the sum of squared observations by $\sum x_i^2$, or simply by $\sum x^2$, and so forth.

Finally, it is important to note that, when discussing descriptive measures, one should distinguish between so-called **ungrouped** data and **grouped** data. Description of ungrouped data entails the study of descriptive measures for the original set of raw data, while grouped data is data that has already been arranged in a frequency table.

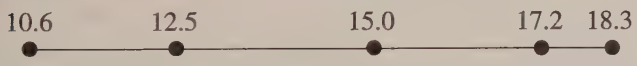
2 Measures of location: ungrouped data

2.1 The arithmetic mean

Consider the following five numbers:

$x_1 = 17.2,\ x_2 = 18.3,\ x_3 = 12.5,\ x_4 = 10.6$ en $x_5 = 15.0$

A scatter diagram gives only a vague indication of the order of magnitude of the numbers:



The best-known single numerical value used to indicate the central position/location of the numbers is the arithmetic mean or average, $\bar{x}$, which, for the above numbers, is calculated as follows:

$$\bar{x} = \frac{17.2 + 18.3 + 12.5 + 10.6 + 15.0}{5}$$

$$= 14.7$$

The **arithmetic mean (or average)** of a set of observations $x_1, x_2, \ldots, x_n$ is defined by

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$= \frac{\sum x}{n}$$

The arithmetic mean of prices per square metre for the simple random sample of 100 residential houses in the Johannesburg area (as arranged in ascending order in Table 3.5) is

$$\bar{x} = \frac{273.21 + 275.76 + \ldots + 1\,663.16}{100}$$

$$= \frac{90\,931.60}{100}$$

$$= 909.32 \text{ rand}$$

The following physical representation can facilitate the interpretation of the arithmetic mean: Suppose that each of the data values is represented by a small weight placed on a thin rod of absolutely homogeneous material in such a way that the distances between the weights correspond exactly to the distances between the data values, and that all the weights have exactly the same mass. The arithmetic mean will then be the balancing point of the rod carrying the weights. Figure 4.5 illustrates this interpretation of the five points.

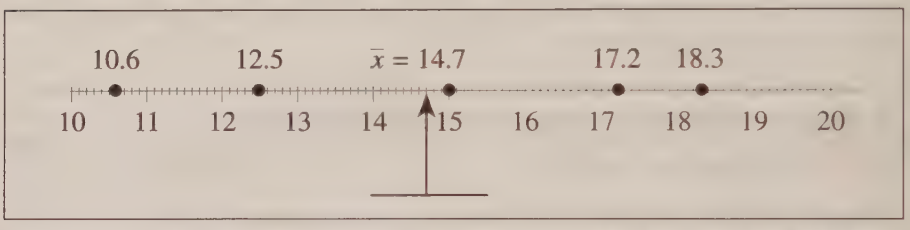


Figure 4.5 Physical interpretation of the arithmetic mean as the balancing point of a data set

Notes

(a) A quick method of computing the arithmetic mean of a set of values that are all more or less the same size, is to subtract the same constant, say a, from each value and to calculate the arithmetic mean of the differences that are obtained. Then add the constant to the arithmetic mean of the differences. In symbols this computation can be represented as follows:

$$\bar{x} = \frac{\Sigma(x - a)}{n} + a$$

(b) The arithmetic mean as discussed above is a descriptive measure of location for a sample of observations. The mean of a population of size N can similarly be defined as

$$\mu = \frac{x_1 + x_2 + \ldots x_N}{N}$$

$$= \frac{\Sigma x}{N}$$

The most important difference between the arithmetic mean, $\bar{x}$, and the population mean, μ, is that $\bar{x}$ varies from one sample to the next, while μ is a fixed value which can be computed only if the entire population has been observed (by means of a census). Chapter 11 will include a discussion of methods showing how μ can be estimated from the value of $\bar{x}$.

EXAMPLE 4.1

The six figures below reflect the maximum monthly temperatures (over a number of years) for the summer months of October to March in Upington (*South African Statistics*, 1990):

39.4 40.9 42.1 41.7 41.4 38.7 degrees Celcius

A convenient way of calculating the average monthly maximum temperature for these months, is to subtract 40.0 from each of the figures. This gives the following values:

−0.6 0.9 2.1 1.7 1.4 −1.3

Therefore:

$$\bar{x} = \frac{-0.6 + 0.9 + 2.1 + 1.7 + 1.4 - 1.3}{6} + 40.0$$

$$= 0.7 + 40.0$$

$$= 40.7 \text{ degrees Celsius}$$

Properties of the arithmetic mean

The arithmetic mean is the best-known and most commonly used measure of location and its precise meaning is easily explained. Any raw data set has a unique arithmetic mean; in other words every data set has only one arithmetic mean. It is obvious from this definition that the arithmetic mean is based on each figure in the data set and can therefore be regarded as a representative value for the entire data set. Although the value of the arithmetic mean usually varies only slightly from one sample to the next (another sample of 100 prices per square metre for houses in Johannesburg would give an arithmetic mean that differs only slightly from 909.32), outlier values significantly affect the value of $\bar{x}$. This effect is more prominent if the sample is small. Take as an example the following ten values of the random sample of 100 prices per square metre (see Chapter 2) namely

809.35	335.05	795.93	1 000.00	1 089.74
526.32	877.86	1 000.00	1 111.11	1 335.23

for which $\bar{x} = 888.06$. If the second value, namely 335.05, were to be replaced by another value from the sample, for example: 639.27, then $\bar{x} = 918.48$, which is much closer to 909.32, the mean of the sample of 100 observations.

Finally, $\bar{x}$ has certain mathematical properties which render it more suitable for further analysis than other measures of location. One such mathematical property emerges clearly when the joint arithmetic mean of different sets of

data is computed from the arithmetic means of the individual sets. This can be done by using the **weighted arithmetic mean**.

2.2 The weighted arithmetic mean

There are frequent press releases on the official level of dams in important catchment areas. If, for example, the average content of dams in the Vaal River catchment area is reported as currently being 70%, how should this mean be interpreted?

EXAMPLE 4.2

At the end of August 1992 (the end of an exceptionally dry period) the net full supply capacity (FSC) in millions of cubic metres and the percentage content of the three most important dams in the Vaal River catchment area were as follows:

Dam	Net FSC	Percentage content
Vaal Dam	2 529	20
Bloemhof Dam	1 269	20
Sterkfontein Dam	2 617	99

The average percentage content of these dams was

$$= \frac{(2\,529 \times 20) + (1\,269 \times 20) + (2\,617 \times 99)}{2\,529 + 1\,269 + 2\,617}$$

$$= \frac{335\,043}{6\,415}$$

$$= 52.2\%$$

Since the three dams are not of equal size (the content capacities differ), the arithmetic mean of the percentage content of the three dams cannot be used to determine the average level of the dams. The value of 52.2% is the weighted mean of the content of the three dams.

Let $y_1, y_2, \ldots, y_n$ denote n numbers with the relative importance of $w_1, w_2, \ldots, w_n$. The **weighted arithmetic mean** of $y_1, y_2, \ldots, y_n$ is defined by

$$\bar{y}_w = \frac{\Sigma wy}{\Sigma w}$$

An important application of this type of mean is found in the computation of the joint arithmetic mean, $\bar{x}_w$, of various sets of data. If n_i and $\bar{x}_i$ respectively indicate the size and arithmetic mean of the ith data set, $i = 1, 2, \ldots, k$, then

$$\bar{x}_w = \frac{\Sigma n_i \bar{x}_i}{\Sigma n_i}$$

EXAMPLE 4.3

Suppose we want to determine the arithmetic mean of the 100 selling prices for houses in the four cities: Durban, Johannesburg, Cape Town and Pretoria (see Table 2.3). The following information is available for the subsamples from the four strata:

Durban:	$n_1 = 18$	$\bar{x}_1 = 143.47$
Johannesburg:	$n_2 = 41$	$\bar{x}_2 = 175.51$
Pretoria:	$n_3 = 21$	$\bar{x}_3 = 149.90$
Cape Town:	$n_4 = 20$	$\bar{x}_4 = 143.29$

The arithmetic mean of the total sample of 100 is therefore:

$$\bar{x}_w = \frac{(18 \times 143.47) + (41 \times 175.51) + (21 \times 149.90) + (20 \times 143.29)}{18 + 41 + 21 + 20}$$

$$= \frac{15\,792.07}{100}$$

$$= 157.29 \text{ thousand rand}$$

2.3 The median

Consider the average monthly rainfall (in millimetres) for East London over the last five months of the year:

41.5 97.4 111.0 94.3 87.5

The data array of these observations in ascending order is:

41.5 87.5 94.3 97.4 111.0

The value in the centre or middle of such an array is known as the median (*me*).

> The **median** of n observations is defined as the $\left(\frac{n+1}{2}\right)$th value in the data array.

For the above five rainfall figures the median is exactly in the middle of the data array, in other words

$me = 94.3$

EXAMPLE 4.4

The average monthly rainfall figures for East London over the twelve months of the year are as follows:

75.7 77.5 99.2 68.8 47.6 35.1
28.9 41.5 97.4 111.0 94.3 87.5

The array of the 12 rainfall figures is as follows:

28.9 35.1 41.5 47.6 68.8 75.7
77.5 87.5 94.3 97.4 99.2 111.0

In this case the median is between the sixth and seventh values of the data array because there is an even number of items, in other words,

$$\begin{aligned} me &= 75.7 + 0.5(77.5 - 75.7) \\ &= 75.7 + 0.9 \\ &= 76.6 \end{aligned}$$

Properties of the median

The median is a simple measure and its meaning is easily understood. It can be calculated from any kind of data set. It is unique and can be used in cases where there is no point in calculating the arithmetic mean. Suppose, for example, that the centre or central position of the following six rainfall figures arranged in ascending order, needs to be determined:

51.4 69.3 72.5 81.1 88.0 169.7

Here $\bar{x} = 88.7$ and $me = 76.8$. This example illustrates that, in contrast with the arithmetic mean, the median is hardly influenced by outliers. If outliers do occur in a data set, the median is therefore a more representative measure of location than the arithmetic mean.

The most important disadvantages of the median are that it is time consuming to calculate from large unarranged data sets and it does not have the useful mathematical properties that the arithmetic mean has. It is, for example, not possible to calculate the median selling price for the four cities, Johannesburg, Pretoria, Durban and Cape Town, from the separate median values obtained from the samples of the four strata.

2.4 The mode

A property common to both the arithmetic mean and the median is that neither necessarily corresponds with any of the values for which it has been calculated. Clothing and footwear manufacturers wishing to determine the most typical size, would not find $\bar{x}$ and me very useful as middle values. This necessitates a further measure of central tendency, namely the mode (mo).

The **mode** is the value most often occurring in the data set.

Suppose that, in a random sample of fifty male clients from a particular shoe shop, the following distribution of shoe sizes is found:

Shoe size	5	6	7	8	9	10	11	12	13
Number of clients	2	4	5	6	8	13	6	4	2

Here $mo = 10$.

Although the mode is not very useful in analysing continuous ungrouped data, it can be calculated theoretically. Thus, for example, the mode of the 41 selling prices of houses in Johannesburg is R65 000 (see Table 2.3).

Properties of the mode

In an ungrouped data set the mode is a simple and easily computable measure of central tendency. Like the median, it is not affected by outliers. The mode is best applied in the analysis of discrete data as is clearly illustrated by the above example of shoe sizes. For nominal data the mode also yields important information. An example would be determining the most popular subject combination in a particular university faculty. A disadvantage attached to the mode is that it is not always uniquely defined. If two or more values share the description 'most popular', the data has two or more modes, in which case one refers to bimodal or multimodal data. Just like the median, the mode does not have the mathematical properties which would make it useful in more advanced statistical analyses.

2.5 The geometric mean

When changes in financial and demographic data have been observed over a period of time an additional measure of central tendency is required. For example, consider a bank that has paid varying rates of interest over a period of five years for a certain type of investment as indicated in the following table:

Year	Interest rate	Growth factor
1	16.75	1.1675
2	16.25	1.1625
3	18.00	1.1800
4	19.50	1.1950
5	19.75	1.1975

The arithmetic mean of the interest rates for the five years is

$$\bar{x} = \frac{16.75 + 16.25 + 18.00 + 19.50 + 19.75}{5}$$

$$= 18.05$$

An investment of R100 000 over a period of five years at this average interest rate (compounded interest) then grows to an amount of

$$100\,000(1.1805)^5 = 229\,260.88 \text{ rand}$$

This amount differs from the amount to which the investment has grown and which is calculated as follows:

$$100\,000(1.1675 \times 1.1625 \times 1.1800 \times 1.1950 \times 1.1975) = 229\,179.25 \text{ rand}$$

The difference of R81.63 between the above calculated end values can be eliminated if the end amount is based on the geometric mean of the growth factors (as given in the third column of the above table).

The **geometric mean** of the values $x_1, x_2, \ldots, x_n$ is defined by

$$gm = \left(x_1 \cdot x_2 \cdot \ldots \cdot x_n\right)^{\left(\frac{1}{n}\right)}$$

On the basis of the above growth factors, the geometric mean can be calculated as follows:

$$gm = \left(1.1675 \times 1.1625 \times 1.1800 \times 1.1950 \times 1.1975\right)^{\left(\frac{1}{5}\right)}$$
$$= 1.180415917$$

From this we can see that, on average, an investment grows by $100(1.180415917 - 1) = 18.0415917\%$ per annum. If we therefore compute the end amount over a period of five years at an average interest rate of 18.0415917%, the following result is obtained:

$$\text{end value} = 100\,000(1.180415917)^5$$
$$= 229\,179.25 \text{ rand}$$

which corresponds exactly with the actual end value as determined above.

Economic and demographic time series data, for example, indices or population numbers, often necessitate the computation of the geometric mean as a measure of location. The following example serves as a further illustration.

EXAMPLE 4.5

The consumer price index of the RSA for the twelve months of 1991 is given in the second column of the following table:

Table for calculating growth rate

Month	Price index	Growth factor	Growth rate
January	107.7	–	–
February	109.2	$\frac{109.2}{107.7} = 1.01393$	1.393
March	109.8	$\frac{109.8}{109.2} = 1.00549$	0.549
April	111.1	$\frac{111.1}{109.8} = 1.01184$	1.184
May	112.9	$\frac{112.9}{111.1} = 1.01620$	1.620
June	114.2	$\frac{114.2}{112.9} = 1.01151$	1.151
July	115.8	$\frac{115.8}{114.2} = 1.01401$	1.401
August	117.2	$\frac{117.2}{115.8} = 1.01209$	1.209
September	118.6	$\frac{118.6}{117.2} = 1.01195$	1.195
October	120.9	$\frac{120.9}{118.6} = 1.01939$	1.939
November	122.1	$\frac{122.1}{120.9} = 1.00993$	0.993
December	123.8	$\frac{123.8}{122.1} = 1.01392$	1.392

The geometric mean of the growth factors is now:

$$gm = (1.01393 \times 1.00549 \times \ldots \times 1.01392)^{\left(\frac{1}{11}\right)}$$
$$= 1.01275$$

This figure means that, during 1991, the South African Consumer Price Index grew by an average of 1.275% per month. An average monthly growth factor of 1.01275 (or a growth rate of 1.275%) over eleven months from the end of January to the end of December 1991 with a starting point of 107.7 yields

$$107.7(1.01275)^{11} = 123.8$$

which corresponds with the December consumer price index in the table.

The above calculations can be simplified considerably by noting that when the growth factors are multiplied, the same results can be obtained by dividing the numerator of the last factor by the denominator of the first factor and by raising this quotient to the power $\frac{1}{11}$:

$$gm = \left(\frac{123.8}{107.7}\right)^{\left(\frac{1}{11}\right)}$$
$$= 1.01275$$

Properties of the geometric mean

The geometric mean is a less known measure that is not as easily understood as the other measures of location, but can be used specifically for data representing growth (economic and demographic data) and for ratios (for example: indices and percentages). If the data comprises a mixture of positive and negative figures or if one or more of the figures are zero, the geometric mean is obviously not a meaningful measure of location.

2.6 The quantiles

All the preceding measures have been defined to describe the midpoint or centre of a data set in some or other way. However, there are also measures of location that aim to reflect a broader picture of the data set. It may, for example, be important to know whether a particular person is in the top 25% of the economically active population. The measures of location used to provide an answer to these and similar questions are known as quantiles. **Quantiles** is a collective noun for the so-called quartiles, deciles and percentiles.

The three **quartiles** q_1, q_2 and q_3 divide a data array into four equal parts, while the nine **deciles** d_1, d_2, ..., d_9 and the 99 **percentiles** $p_1, p_2, \ldots, p_{99}$ respectively divide the array into 10 and 100 equal parts.

Figure 4.6 gives a schematic representation of the similarities and the relations between the different quantiles. From this figure it becomes evident that every quantile can be regarded as a percentile, so that the calculation for every quantile can be given in terms of the ith percentile.

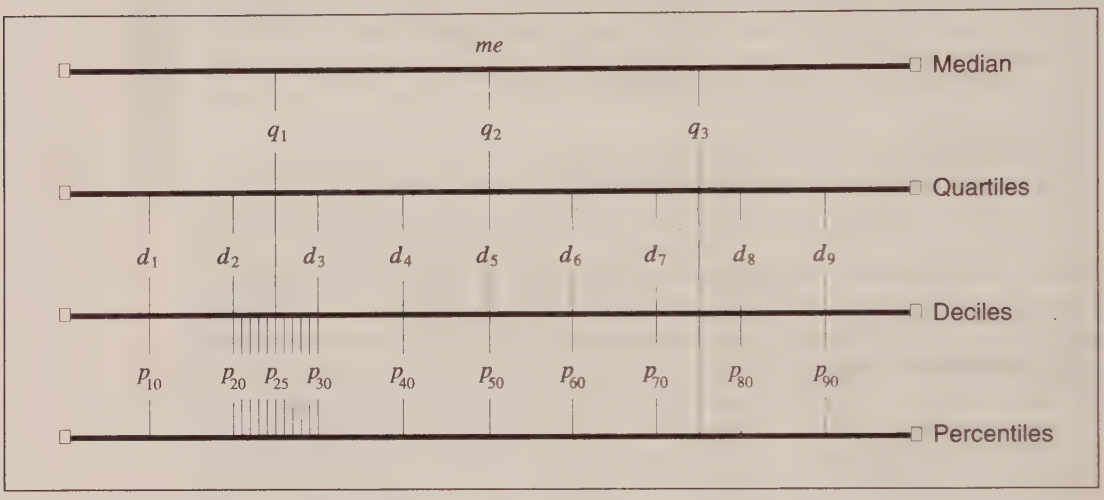


Figure 4.6 Relation between quantiles

If $1 \leq \dfrac{i(n+1)}{100} \leq n$, the ith percentile can be determined as the $\dfrac{i(n+1)}{100}$ th value in the data array.

EXAMPLE 4.6

Consider the average number of rainy days for the months January to December in Durban (as observed over a number of years):

14 12 12 9 5 4 4 7 9 14 15 15

The array of the twelve values is

4 4 5 7 9 9 12 12 14 14 15 15

To illustrate the above calculation formula q_1 and d_9 will now be calculated.

q_1 = value of $\dfrac{25(12+1)}{100}$ th observation in the data array

= value of 3.25th observation

= 3rd observation + 0.25 (difference between 3rd and 4th observation)

Therefore

$q_1 = 5 + 0.25(7 - 5)$

$= 5.5$ days

$$d_9 = \text{value of } \frac{90(12+1)}{100} \text{ th observation in the data array}$$
$$= \text{value of } 11.7\text{th observation}$$
$$= 11\text{th observation} + 0.7\,(\text{difference between 11th and 12th observation})$$

Therefore

$$d_9 = 15 + 0.7(15 - 15)$$
$$= 15 \text{ days}$$

Properties of the quantiles

There is no point in calculating deciles and percentiles for small data sets. However, these quantiles are very useful for describing larger data sets. Since the arrangement of a large data set in ascending order is time consuming, it is better to group the data set in a frequency table prior to calculating the quantiles. This matter is discussed in Section 3. Quantiles are fairly insensitive to outliers although there are exceptions: If the last value of twelve rainy days, as set out in Example 4.6, had been, for example, twenty days, most of the deciles would have remained the same, but d_9 would have been considerably larger. The calculation of a set of quantiles has the advantage that it provides far more information than a single central value would have provided.

3 Measures of location: grouped data

The advantages attached to grouping a data set in a frequency table have already been pointed out in Chapter 3. Moreover, descriptive measures in large data sets can be calculated with more ease if the data is grouped. Once the data has been grouped for calculating descriptive measures it is usually assumed that the original set of raw or ungrouped data is no longer available. This fact means that a part of the information contained in the original data is lost.

To be able to define descriptive measures meaningfully, the following assumption is made regarding the distribution of the values within each class interval:

> **Assumption for grouped data**: The values within each class interval are distributed uniformly over the entire class and consequently all the values within a class can be represented by the class mark, or midpoint.

This assumption usually means that the values of the descriptive measures differ somewhat from the values that would have been obtained from the corresponding ungrouped data set. However, the inaccuracy is usually excusable if the advantages attached to this technique are taken into account. In fact, data is often available only in the form of a frequency table. In such a case the above assumption is essential if descriptive measures are to be calculated.

3.1 The arithmetic mean

Consider a grouped set of n values that are divided into k classes with class marks $x_1, x_2, \ldots, x_k$ and class frequencies $f_1, f_2, \ldots, f_k$ where $\Sigma f = n$. In order to define the arithmetic mean, the assumption for grouped data is used, in other words we assume that f_i values are equal to x_i, $i = 1, 2, \ldots, k$.

The **arithmetic mean for grouped data** is defined as the weighted arithmetic mean of the class marks, namely

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots + f_kx_k}{n}$$

$$= \frac{\Sigma fx}{n}$$

EXAMPLE 4.7

Table 4.1 is used as an aid for the calculation of the arithmetic mean of the 100 prices per square metre for houses in Johannesburg, as set out in Table 3.7.

From Table 4.1 it follows that $\Sigma fx = 91\,400$, so that

$$\bar{x} = \frac{91\,400}{100}$$

$$= 914.00 \text{ rand}$$

This value compares well with the true arithmetic mean $\bar{x} = 909.32$ rand as calculated in Section 2.1 for the ungrouped data set.

Table 4.1 Auxiliary table for calculating the arithmetic mean of the 100 prices per square metre as contained in Table 3.7

Class interval	Class mark x_i	Class frequency f_i	f_ix_i
[100, 300)	200	2	400
[300, 500)	400	8	3 200
[500, 700)	600	18	10 800
[700, 900)	800	25	20 000
[900, 1 100)	1 000	19	19 000
[1 100, 1 300)	1 200	12	14 400
[1 300, 1 500)	1 400	10	14 000
[1 500, 1 700)	1 600	6	9 600
		100	91 400

Note

There are various methods of simplifying the calculation of the arithmetic mean of grouped data. Thus, for instance, one can first subtract a constant value, for example, a, from each class mark and the arithmetic mean can then be calculated from the formula

$$\bar{x} = \frac{\Sigma f(x - a)}{n} + a$$

These methods are only of interest when a pocket calculator is not available.

3.2 The quantiles

Just as in the case of ungrouped data, the quantiles divide the data into 2, 4, 10 or 100 equal parts, vertical lines through the quantiles divide the area of the histogram of a grouped data set into 2, 4, 10 or 100 equal parts.

> The ***i* th percentile** can be defined as the value to the left of which $i\%$ of the total area of the histogram lies.

Figure 4.7 shows the fourth decile or fortieth percentile of the 100 prices per square metre of houses in Johannesburg (see Table 3.7). A total of 40% of the total area of the histogram lies to the left of the vertical line through d_4. It becomes apparent from this figure that, when calculating the quantiles in the grouped data, the number of values in each class is distributed uniformly over the interval (the assumption for grouped data). The calculation of the quantiles can be done graphically by means of the cumulative frequency polygon (see Example 4.8).

> The ***i* th percentile** is the value that corresponds with the $\frac{in}{100}$ th value in the cumulative frequency distribution.

The percentile number i need not be an integer. It is, for example, often important to calculate the 33.3th percentile. Consider the following example:

EXAMPLE 4.8

The data contained in Table 4.2 represents the ages of the 10 691 South Africans who died during 1988 in road accidents. The cumulative frequency distribution is shown in the last two columns of the table and the cumulative frequency polygon in Figure 4.8.

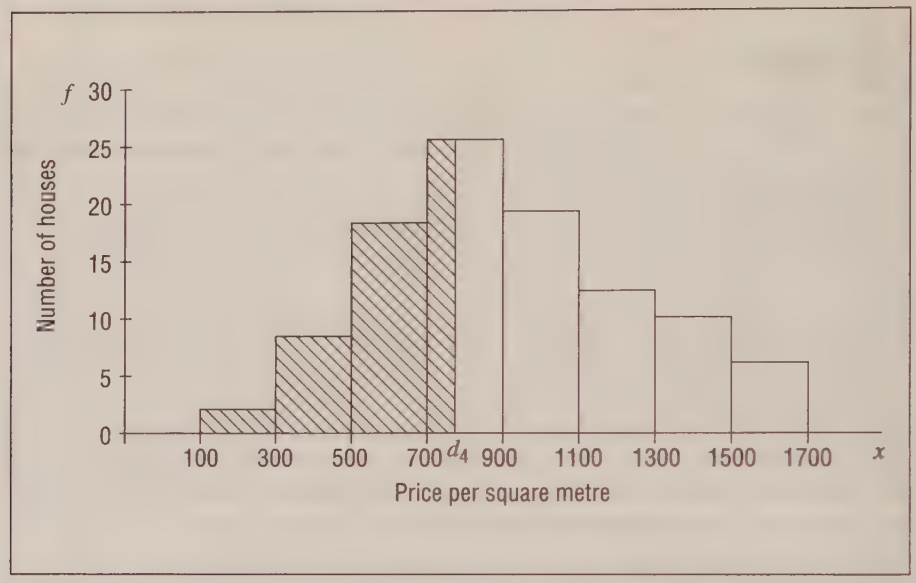


Figure 4.7 Interpretation of d_4 or p_{40} in the case of grouped data

Table 4.2 Ages of South Africans killed in road accidents during 1988

Age	Frequency	Class boundary	Cumulative frequency
(0, 5]	358	5	358
(5, 10]	497	10	855
(10, 15]	382	15	1 237
(15, 20]	675	20	1 912
(20, 25]	1 306	25	3 218
(25, 30]	1 655	30	4 873
(30, 35]	1 478	35	6 351
(35, 40]	1 299	40	7 650
(40, 45]	1 037	45	8 687
(45, 50]	589	50	9 276
(50, 55]	519	55	9 795
(55, 60]	259	60	10 054
60+	637	–	10 691

Source: *SA Statistics*, 1990, *Central Statistical Service*

Figure 4.8 demonstrates how the *me* and $p_{33.3}$ for a grouped data set can be read from the cumulative frequency polygon. The values of *me* and $p_{33.3}$ are read on the *x*-axis and correspond respectively with 5 345.5 and 3 563.67 on the *F*-axis. The following is found:

$me \approx 32$ years
$p_{33.3} \approx 26$ years

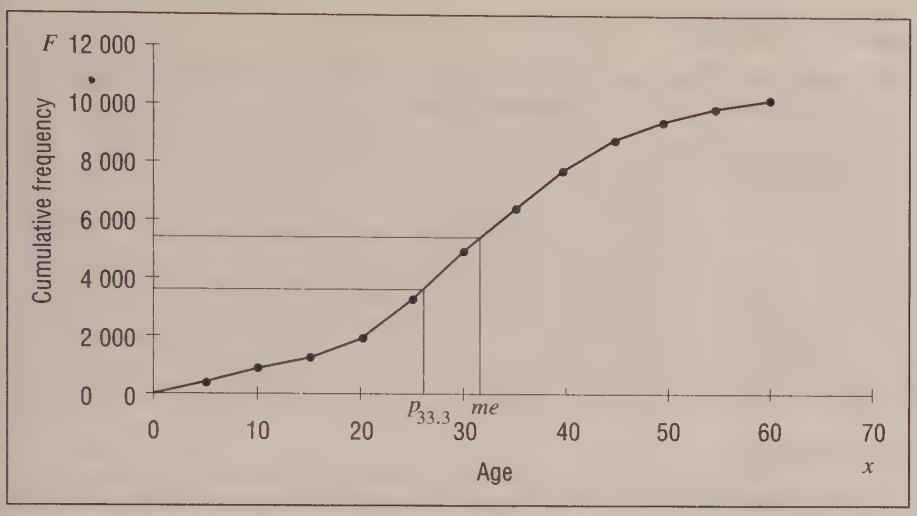


Figure 4.8 Graphic determination of *me* and $p_{33.3}$ for ages of persons killed in road accidents during 1988

Note

Note that in respect of grouped data the position of p_i is given by $\frac{in}{100}$, in contrast with ungrouped data where the position is given by $\frac{i(n+1)}{100}$. This difference follows from the fact that we assume in the case of grouped data that the data is spread continuously over the range.

A more accurate value of the quantiles can be found by means of an analytic method, using the information contained in the cumulative frequency distribution and the assumption for grouped data from the following formula:

$$p_i = k_l + \frac{(k_u - k_l)\left(\frac{in}{100} - F_{k_l}\right)}{f_{p_i}}$$

where p_i = ith percentile

k_l = lower class boundary of the interval containing p_i

k_u = upper class boundary of the interval containing p_i

F_{k_l} = cumulative frequency smaller than k_l

f_{p_i} = frequency of interval containing p_i

As an illustration of the use of this formula *me* and $p_{33.3}$ will now be calculated for the data set of Example 4.8.

EXAMPLE 4.9

The median and 33.3rd percentile of the ages of victims of road accidents during 1988, calculated according to the above formula, are as follows:

$$me = 30.00 + \frac{5(5\,345.5 - 4\,873)}{1\,478}$$

$$= 31.60$$

$$p_{33.3} = 25.00 + \frac{5(3\,560.103 - 3\,218)}{1\,655}$$

$$= 26.03$$

3.3 The mode

It is difficult to determine the mode of a grouped data set accurately. As in the case of ungrouped data the interval with the highest frequency can be regarded as an indication of the position of the mode. This class interval is called the modal interval. The mode is obtained from the histogram by means of a graphic method as explained in Figure 4.9. The histogram in this figure represents the frequency distribution of the ages of victims of road accidents during 1988 (on account of the open interval only the distribution of persons under the age of sixty is represented in the histogram). The modal interval is (25; 30) and the mode can be found graphically as the point where the perpendicular through the point of intersection of AB and CD meets the x-axis. The following value for the mode is found by means of this technique:

$mo \approx 28$ years

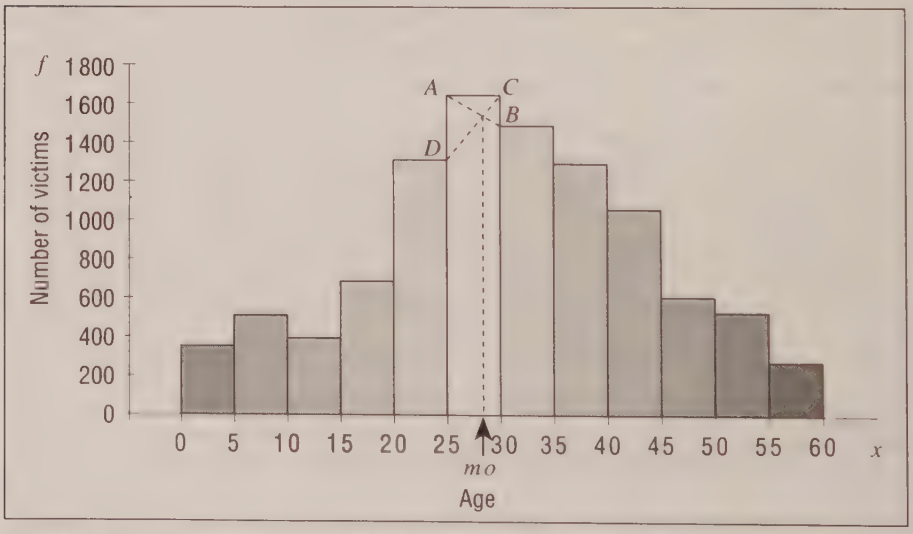


Figure 4.9 Graphic determination of mo for the ages of victims of road accidents

The following formula that is based on the above graphic technique can be used to determine the value of the mode more accurately:

$$mo = k_l + \frac{(k_u - k_l)(f_m - f_{m-1})}{(2f_m - f_{m-1} - f_{m+1})}$$

where k_l = lower class boundary of modal interval
k_u = upper class boundary of modal interval
f_m = frequency of modal interval
f_{m-1} = frequency of interval preceding modal interval
f_{m+1} = frequency of interval following modal interval.

By using the above formula, the mode for the data set in Table 4.2 (as represented graphically in Figure 4.9), is calculated as follows:

$$mo = 25 + \frac{5(1\,655 - 1\,306)}{2 \times 1\,655 - 1\,306 - 1\,478}$$

$$= 28.32 \text{ years}$$

4 Concluding remarks

The properties of the different measures of location were discussed in the preceding sections of this chapter. One can add that, in the case of a grouped data set containing one or more open-ended class intervals, it is impossible to calculate the arithmetic mean. On the other hand, one can calculate the median and mode in such a case. The choice of a specific set of class boundaries affects the mode to a larger extent than it does the median and the arithmetic mean.

Each of the measures of location discussed in this chapter describes a data set in a specific way. For a given data set one should note the nature of the data and the purpose of the analysis before deciding on the measure that will best describe the data. Since the analysis of a data set is not usually confined to a study of the properties of location, but also involves further statistical analyses within the broad framework of descriptive statistics and statistical inference, the arithmetic mean is the most commonly used and most important measure of location.

Finally, it is important to mention the fact that the differences between some measures of location can be used to describe other properties, such as spread and symmetry. This will be discussed in more detail in Chapter 5.

Exercises

1 Take the first ten houses in each of the four cities, namely Durban, Johannesburg, Cape Town and Pretoria as random samples of size 10 each (Appendix A, Data Set 1).

(a) Calculate the arithmetic mean and the median for the selling prices (variable: price) of each of these samples.

(b) Calculate the average selling price for the combined sample of 40 houses by using the arithmetic means determined in (a). Can the median selling price of the forty houses be determined in the same way?

(c) Calculate the mode of the plot sizes (variable: size) for each of the four samples, and also for the combined sample.

2 The net full supply capacity (FSC)(in millions of cubic metres) in the various regions and catchment areas in South Africa, and also the percentage content as on 31 August 1992 are given in the table below:

Region/catchment area	FSC	Percentage capacity
Western Transvaal	953	0,34
Eastern Transvaal	2 516	0,42
Vaal River	7 474	0,53
Upper Orange River	9 655	0,38
Lower Orange River	21	1,00
Western Cape	1 672	0,78
Eastern Cape	1 042	0,37
Natal	4 686	0,52

(a) Calculate the arithmetic mean and the median of the full supply capacity for the eight regions. Which of the measures best reflects the locality feature?

(b) Calculate the arithmetic mean and the weighted arithmetic mean for the percentage content of the eight regions. Which measure is the better of the two in this case?

3 The number of telephone users in the RSA for the period 1987 to 1989 has been tabled according to type. This information is reflected in the accompanying table. Determine the modal type of user for each year.

Year	Type of users		
	Business	Home	Public call offices
1987	2 241 268	1 957 151	37 503
1988	2 379 051	2 044 548	38 308
1989	2 402 923	2 300 462	40 968

Source: *SA Statistics* 1990, Central Statistical Service

4 The number of nickel and bronze coins minted by the South African Mint for the period 1987 to 1989 are classified by denomination in the following table:

Denomination	1987	1988	1989
Nickel			
R2			51 805 789
R1	12 189 315	21 401 353	13 887 930
50c	19 085 623	27 711 032	17 269 121
20c	36 918 423	43 128 032	61 182 721
10c	43 248 423	48 580 030	62 206 721
5c	42 800 423	111 177 046	9 122 261
Copper			
2c	104 995 423	182 049 030	212 826 721
1c	120 690 423	240 285 082	329 614 721

Source: *SA Statistics* 1990, Central Statistical Service

What is the modal denomination for each of the years if
(a) only nickel coins; or
(b) all coins
are taken into consideration?

5 Use the properties of the summation notation to prove that

(a) $\bar{x} = \dfrac{\sum (x - a)}{n} + a$

(b) $\bar{x}_w = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

where $\bar{x}_1$ and $\bar{x}_2$ are the respective averages of two samples of sizes n_1 and n_2 and $\bar{x}_w = \dfrac{\text{sum of the values of the two samples}}{\text{total number of observations}}$

6 According to the publication *Education Realities in South Africa 1990* the number of students enrolled (as on 10 March 1990) and the percentage of students who passed at seventeen South African universities are given in the table below. Calculate:
(a) the average number of students per university;
(b) the median number of students for the seventeen universities; and
(c) the weighted arithmetic mean of the percentage of students who passed during 1989, using the number of students enrolled in 1990 as weights.

University	Number of students enrolled (10 March 1990)	Percentage passed (1989)
Cape Town	13 545	82
Durban-Westville	7 485	64
Medical University of Southern Africa	1 983	72
Natal	12 803	82
North	9 774	65
Orange Free State	9 242	75
Port Elizabeth	4 768	77
PU for CHE	8 658	75
Pretoria	22 976	76
Rand Afrikaans University	9 055	75
Rhodes	3 802	80
UNISA*	108 016	47
Stellenbosch	13 830	81
Western Cape	12 732	66
Witwatersrand	18 233	80
Zululand	5 195	67
Vista	24 818	63

* Mainly distance teaching

7 Chapter 5, Table 5.11 contains information on the number of beds available, bed occupancy rates and the financial turnover for February and July for a specific year for the 22 rest camps in the Kruger National Park. Determine:
(a) the arithmetic mean and the median for the number of beds available for July;
(b) the median values of the turnover for each month; and
(c) the weighted arithmetic mean for the July turnover values (use the number of beds available as weights).

8 The South African Consumer Price Index for the twelve months of a number of years is set out in Table 8.5 (Chapter 8). Calculate the growth factors (see Section 2.5) for each of the months from January 1989 to December 1989 and determine the geometric mean of these factors.

9 Repeat Exercise 8 for the twelve months of 1992.

10 The following table reflects student and lecturer numbers at universities and teaching colleges for 1970, 1980 and 1990.

	1970	1980	1990
Students	101 631	184 830	336 994
Lecturers	7 900	10 204	14 029

Source: *RSA Statistics in Brief* 1991, Central Statistical Service

Determine:
(a) the student/lecturer proportion for each of the three years; and
(b) the geometric mean for the annual rate of increase in student numbers, lecturer numbers and the student/lecturer proportion.

11 The economically active population (in thousands) for the RSA (excluding TBVC states) as on 30 June for 1970, 1980 and 1990 was as follows:

1970	1980	1990
7 665	8 732	11 073

Source: *RSA Statistics in Brief* 1991, Central Statistical Service

Calculate the average annual growth rate over the twenty-year period.

12 The number of motor vehicles (in thousands) in the RSA as at 30 June for 1970, 1980 and 1990 was as follows:

1970	1980	1990
2 287	3 794	5 403

Source: *RSA Statistics in Brief* 1991, Central Statistical Service

Calculate the average annual growth rate over the twenty-year period.

13 The Gross Domestic Product (in million rand) for the twenty-year period 1970 to 1990 is reflected in the data below:

1970	1980	1990
75 891	105 121	119 861

Source: *RSA Statistics in Brief* 1991, Central Statistical Service

Calculate the average annual growth rate over this period.

14 Use the data provided in Exercise 2 to calculate the first and third quartiles of the net full supply capacity of the dams in the eight regions.

15 Refer to the numbers of students enrolled as reflected in Exercise 6 to determine:
(a) the universities at which student numbers are higher than the third quartile; and
(b) the first and ninth deciles. Would it be fair to say that 90% of the student numbers at South African universities is evenly spread over the interval with an upper boundary of d_9?

16 Consider the data on the Kruger National Park given in Chapter 5, Table 5.11.
(a) Compile an array of the 22 camps according to

(i) beds available (February); and
(ii) bed occupancy rate (February).

(b) Determine the first quartile and the 33rd percentile of bed occupancy rates for February.

(c) Determine the eighth decile for the beds available in February.

17 The number of pupils (in thousands) from a certain section of the population enrolled at private schools and divided into age groups was as follows on 1 March 1989:

Age	Number	Age	Number
<6	60	13	2 813
6	2 124	14	2 415
7	2 232	15	2 224
8	2 502	16	1 754
9	2 392	17	1 254
10	2 531	18	876
11	2 376	19+	1 184
12	2 694		

Source: *South African Statistics* 1990, Central Statistical Service

Calculate the arithmetic mean, the median and the mode of the pupils' ages. Comment on the meaning, the advantages and the disadvantages of each of the measures for this data.

18 For the frequency distribution of chrome quantities on the steel plates analysed in Chapter 3, Table 3.3 calculate:

(a) the arithmetic mean;
(b) the median;
(c) the first and third quartiles; and
(d) the mode.

19 With reference to the frequency distribution in Chapter 3, Table 3.13, determine the modal number of times per week that respondents have fresh vegetables if

(a) all 528 respondents; and
(b) only those respondents who do not have fresh vegetables every day are taken into account.

20 Consider the two-way frequency distribution of the ages of employees in the mining industry as reflected in Chapter 3, Table 3.18. Calculate, for the coal mining industry:

(a) the average age;
(b) the modal age;
(c) the median age; and
(d) the 95th percentile.

21 Repeat the calculations in Exercise 20 for the gold mining industry.

22 The taxable income and the number of tax returns issued for a specific section of the population for the 1988 tax year are shown in the frequency table below:

Taxable income	Frequency	Taxable income	Frequency
Loss	230	(45 000, 50 000]	1 006
(0, 5 000]	34 369	(50 000, 60 000]	834
(5 000, 10 000]	68 474	(60 000, 70 000]	267
(10 000, 15 000]	63 127	(70 000, 80 000]	78
(15 000, 20 000]	34 560	(80 000, 90 000]	40
(20 000, 25 000]	19 415	(90 000, 100 000]	32
(25 000, 30 000]	10 245	(100 000, 150 000]	48
(30 000, 35 000]	5 590	(150 000, 200 000]	7
(35 000, 40 000]	3 010	(200 000, 250 000]	1
(40 000, 45 000]	1 714	250 000+	3

Source: *South African Statistics* 1990, Central Statistical Service

Calculate for the taxable income:

(a) the median;
(b) the 75th percentile;
(c) the average taxable income of the taxpayers who did not show a loss; and
(d) the mode of taxable income.

23 According to the 1991 census in the RSA the income distribution of three job categories was as shown in the table below:

Summary of results before adjustment for undercount

Annual income (rand)	Engineer, engineering technician, architect and related occupations	Occupations in the natural sciences	Medical, dental and related health services
No income	579	381	1 629
[1, 1 000)	163	98	1 441
[1 000, 3 000)	910	668	4 908
[3 000, 5 000)	749	525	2 493
[5 000, 7 000)	531	341	5 481
[7 000, 10 000)	1 301	661	14 877
[10 000, 30 000)	12 874	8 609	67 515
[30 000, 50 000)	21 509	10 802	22 123
[50 000, 70 000)	16 666	6 842	8 322
[70 000, 100 000)	13 555	5 250	6 478
[100 000, 300 000)	9 412	2 615	7 373
300 000 +	294	66	957

Source: *Population Census* 1991, Central Statistical Service

Calculate:

(a) the median income for each of the occupational groups;
(b) the modal income for each of the occupation groups; and

(c) the income values between which the middle 50% of practitioners in medical and related fields lie.

(d) If we assume that those in the highest income category (R300 000+) earn an average of R350 000 per annum, calculate the average income of each of the three occupational groups.

24 The age distribution of the unemployed in two specific population groups in South Africa in January 1991 was as shown in the table below (frequencies in thousands):

Age	Frequency (group 1)	Frequency (group 2)
< 20	27	5
$[20, 35)$	53	8
≥ 35	17	3

Source: *Demographic Statistics* 1992, Central Statistical Service

Determine for the ages of both groups

(a) the median; and

(b) the mode.

25 Consider the variable amount of beer produced as given in Appendix A, Data Set 5.

(a) Calculate the following for the two fiscal years:

(i) the mean;

(ii) the median; and

(iii) the mode.

(b) Compare the measures for the two years under consideration.

26 Consider the amount of electricity consumed by the brewery in Appendix A, Data Set 5 and repeat Exercise 25.

27 Consider the amount of electricity used to produce one hectolitre of beer (kWh/hℓ) as given in Appendix A, Data Set 5, round the figure off to one decimal and repeat Exercise 25 for this data.

CHAPTER 5

Measures of spread, symmetry and kurtosis

1 Introduction

Chapter 4 highlighted the fact that the frequency distribution of a data set can be described mainly by four characteristics of shape, namely location, spread, symmetry and kurtosis. Measures of location were also discussed and mention was made of the fact that each location measure describes a data set in its own special way. The other three characteristics of shape can be similarly described by appropriate measures. Chapter 5 first discusses measures of spread, and then of symmetry and kurtosis. The chapter is concluded with a brief introduction to exploratory data analysis.

The data below was obtained from the South African Weather Bureau. Suppose the average rainfall figures for the summer months of November to March for Pretoria and Durban are compared.

Table 5.1 Rainfall figures (millimetres) for Pretoria and Durban

Month	Nov	Dec	Jan	Feb	March
Pretoria	119.6	119.9	134.4	113.0	94.2
Durban	110.9	106.6	110.1	124.7	105.7

Simple calculations yield:

Pretoria: $\bar{x} = 116.2$; $me = 119.6$
Durban: $\bar{x} = 111.6$; $me = 110.1$

According to these measures there is apparently a strong similarity between the monthly averages for the two cities. If the above two data sets are arrayed and represented graphically as in the figure below, we gain more insight into the variation in each data set.

For the Pretoria data set the difference between the highest and the lowest monthly averages is

$134.4 - 94.2 = 40.2$ mm

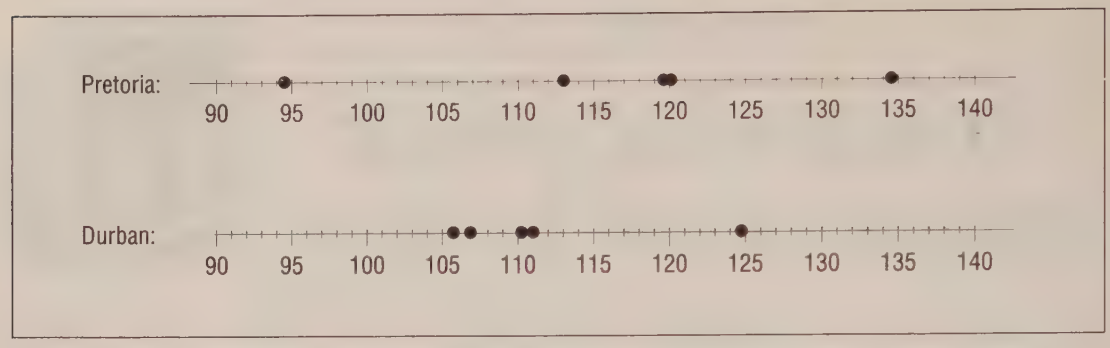


Figure 5.1 Comparison of rainfall figures for Pretoria and Durban for the months November to March

while for the Durban data set it is

$$124.7 - 105.7 = 19.0 \text{ mm}$$

There thus appears to be a greater variation in the monthly average rainfall figures for Pretoria than in those for Durban.

The calculation of measures of spread is important for two main reasons:

(a) Different data sets often have to be compared as in the example above.

(b) Variation is inherent in any data set that is analysed statistically, and information with regard to this characteristic is essential before meaningful conclusions can be drawn from the analysis of a data set. Thus, for example, it is intuitively clear that more value is attached to the arithmetic mean for Durban than to that for Pretoria if estimates of the mean monthly averages for the two cities are required.

2 Measures of spread: ungrouped data

2.1 The range

The **range**, r, is the simplest measure of spread and has already been illustrated in the above example of rainfall data.

The **range** is defined by $r = x_{max} - x_{min}$

The range is particularly applicable in quality control as illustrated in the next example.

EXAMPLE 5.1

Soft drink tins to contain 340 ml each are filled automatically in soft drink factories. Every hour, for example, a random sample of five tins is selected and the

content (in ml) of each tin is determined. The data in Table 5.2 reflects the content of twenty random samples. The last two columns indicate the arithmetic mean, $\bar{x}$, and the range, r, per sample.

Table 5.2 Content of 340 ml soft drink tins for 20 samples of size 5 each

Sample number						$\bar{x}$	r
1	324	337	338	331	355	337.0	31
2	337	336	345	334	332	336.8	13
3	323	333	336	322	338	330.4	16
4	341	325	339	345	324	334.8	21
5	338	336	329	329	340	334.4	11
6	339	334	328	320	351	334.4	31
7	359	356	323	334	347	343.8	36
8	339	336	331	321	335	332.4	18
9	325	341	331	316	331	328.8	25
10	334	324	333	323	322	327.2	12
11	344	345	325	333	333	336.0	20
12	328	343	330	326	355	336.4	29
13	342	323	340	342	324	334.2	19
14	335	344	344	322	339	336.8	22
15	345	326	333	345	325	334.8	20
16	335	345	334	339	348	340.2	14
17	339	321	339	323	329	330.2	18
18	316	316	335	338	325	326.0	22
19	328	332	318	344	354	335.2	36
20	348	351	331	327	323	336.0	28

In order to determine whether the manufacturing process is under **statistical control**, the arithmetic means and ranges are usually plotted on so-called $\bar{x}$ and R **control charts**, as indicated in Figures 5.2(a) and (b).

In addition to the horizontal axis on which the sample number is indicated, a control chart usually contains three or more horizontal lines, including

- the **process average** ((i) on the $\bar{x}$ control chart); and
- **lower and upper action lines** ((ii) on the control charts).

When a plotted point on the control chart lies below the lower or above the upper action line, it is virtually certain that the process was out of control during this sample and that possible causes should be investigated immediately.

From Figure 5.2 it is clear that in the case of sample number 18, $\bar{x}$ is below the lower action line, and consequently the process was not under statistical control during this sample. With respect to the range, however, the process was under statistical control at all times.

2.1.1 Properties of the range

Although the range is the simplest measure of spread and the easiest to calculate, it is based on only the smallest and the largest values and does not

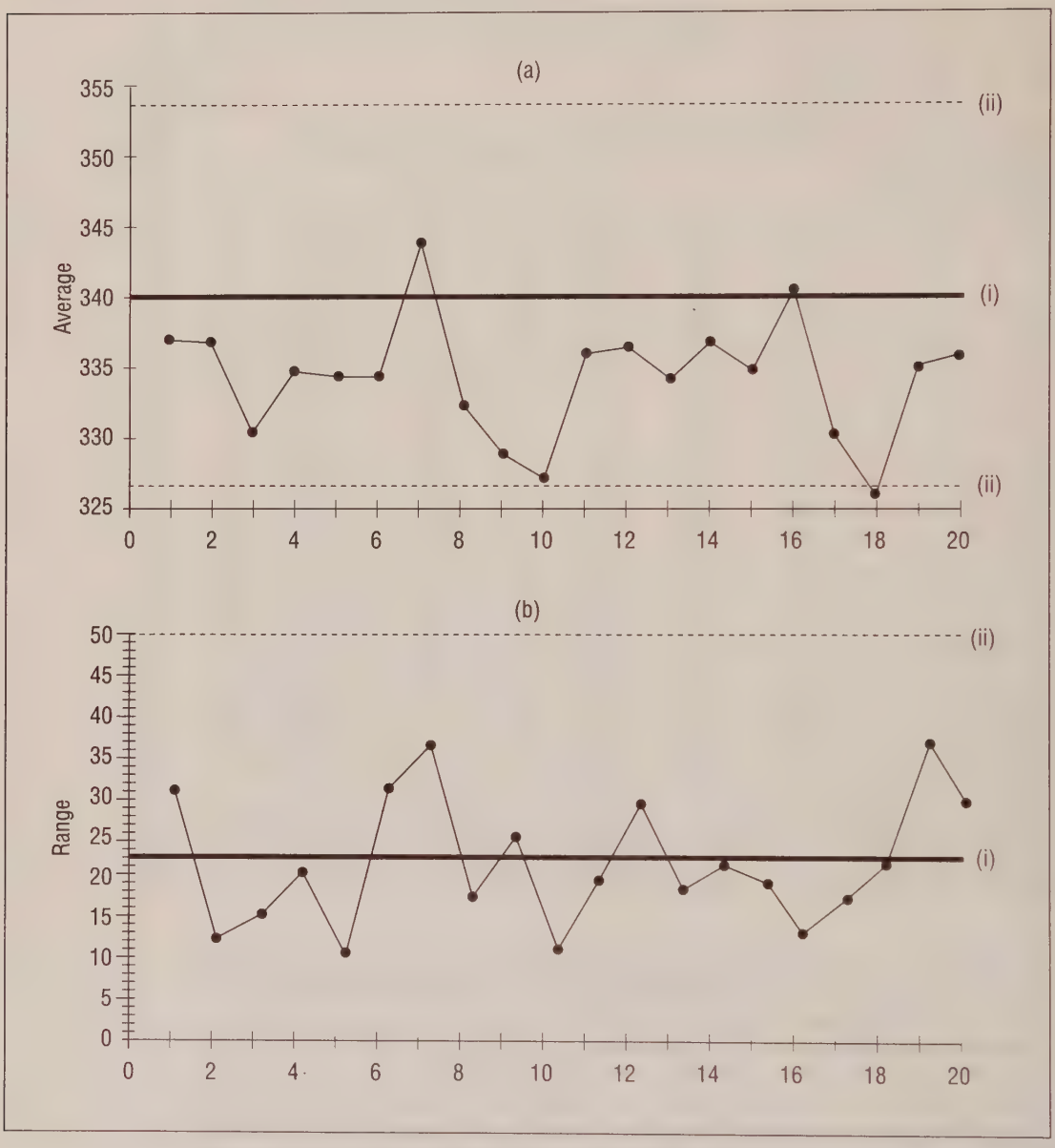


Figure 5.2 $\bar{x}$ and R control charts of soft drink content

therefore reflect the variation of the other values in the data set. Two data sets with the same maximum and minimum values will have exactly the same range, regardless of the spread of the other values between these values. The range is also a very unstable measure. For repeated samples of the same size from a population it fluctuates considerably, as illustrated in Figure 5.2(b).

Note

The determination of the action lines is based on the theory of estimation that is discussed in Chapter 11.

Despite its disadvantages, the range is still important in cases where a measure of spread has to be calculated speedily. In quality control, for example, small samples are drawn on an ongoing basis and analysed immediately. Thus the range of each sample highlights the quality of the items produced. As was indicated in Chapter 3, the range is also useful when compiling a frequency table.

2.2 The quartile deviation

The disadvantages of using the largest and smallest values of a data set for measuring spread can be eliminated to a great extent by defining a measure that is based on a midgroup of values. The middle 50% of the values fall between q_1 and q_3, and logical measures of spread would therefore be the difference $q_3 - q_1$ (called the **interquartile range**), and the mean $\frac{(q_3 - q_2) + (q_2 - q_1)}{2}$ (called the **quartile deviation**).

Formally, the **quartile deviation** is defined by

$$q_d = \frac{(q_3 - q_1)}{2}$$

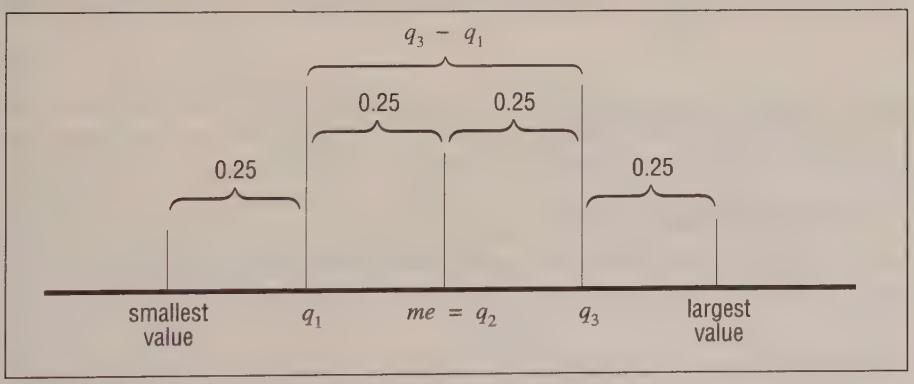


Figure 5.3 The interquartile range

EXAMPLE 5.2

The data below represents the total turnover (in rand) of a certain steakhouse in Pretoria for fourteen so-called 'special' days (holidays, weekends, etc.) in April 1991 (Appendix A, Data Set 4).

2 254.87	1 552.86	2 470.77
1 510.02	1 158.43	1 362.57
1 183.64	1 067.83	1 043.66
830.02	2 305.88	2 886.32
2 989.03	2 537.53	

The quartile deviation of these fourteen values is calculated as follows:
First the data is arrayed, namely

830.02	1 043.66	1 067.83
1 158.43	1 183.64	1 362.57
1 510.02	1 552.86	2 254.87
2 305.88	2 470.77	2 537.53
2 886.32	2 989.03	

$$q_1 = \text{value of the } \frac{25(14+1)}{100} \text{ th observation}$$
$$= \text{value of the } 3.75\text{th observation}.$$

$$\text{Therefore } q_1 = 1\,067.83 + 0.75(1\,158.43 - 1\,067.83)$$
$$= 1\,135.78$$

$$q_3 = \text{value of the } \frac{75(14+1)}{100} \text{ th observation}$$
$$= \text{value of the } 11.25\text{th observation}.$$

$$\text{Therefore } q_3 = 2\,470.77 + 0.25(2\,537.53 - 2\,470.77)$$
$$= 2\,487.46$$

$$\text{The interquartile range is therefore} = 2\,487.46 - 1\,135.78$$
$$= 1\,351.68$$
$$\text{and the quartile deviation } q_d = \frac{2\,487.46 - 1\,135.78}{2}$$
$$= 675.84$$

2.2.1 Properties of the quartile deviation

The quartile deviation is a simple measure of spread that is relatively easy to calculate and is not influenced by outliers. Although this measure is better than the range, it is also based on only some of the observations. For this reason, it is also not suitable for mathematical manipulation with a view to further analysis of the data.

2.3 Standard deviation

With both the range and the quartile deviation the spread of a data set is measured in terms of the difference between only two values, and the variation of the other individual values play no part. If we wish to define a measure of spread that takes all the individual observations into account, the deviation of each value from a point of reference has to be calculated. Because the arithmetic mean is a representative value that is based on all the observations, we use it as the point of reference.

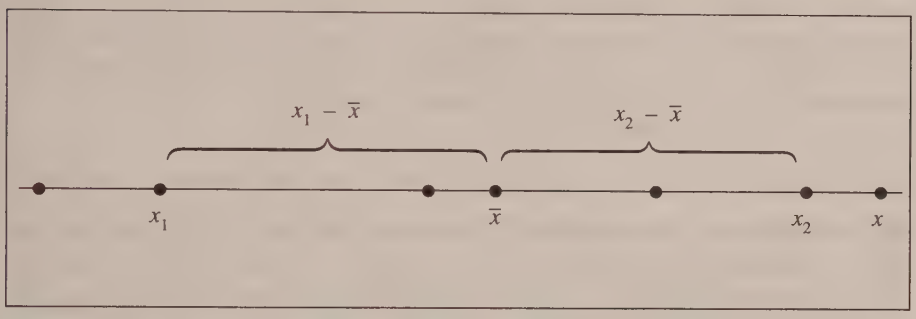


Figure 5.4 The use of $\bar{x}$ as the point of reference for measures of spread

From the above figure it follows that the mean deviation of the values from the arithmetic mean, namely $\frac{\sum(x - \bar{x})}{n}$, is a logical selection for a measure of spread. Table 5.3 shows the calculation of this measure for the rainfall data for Pretoria and Durban as reflected in Table 5.1. As can be seen from this table,

$$\frac{\sum(x - \bar{x})}{n} = 0$$

for both data sets. On closer inspection this appears to be a general result, because some values have a positive and others a negative deviation from $\bar{x}$ and because $\bar{x}$ is in fact the centre of gravity of the observations. When measuring spread, the directions of the signs of the deviations are not important and only the numerical values of the differences should be taken into account. These are obtained by taking the absolute value of each deviation and calculating the mean of the absolute deviations. The measure thus obtained is called the mean absolute deviation.

The **mean absolute deviation** is defined by

$$ma = \frac{\sum|x - \bar{x}|}{n}$$

The mean absolute deviations of the rainfall data in Table 5.1 are also calculated in Table 5.3. The difference in spread between Pretoria and Durban with respect to monthly average rainfall is clearly indicated by these calculated values.

Table 5.3 Calculation of $\frac{\sum(x-\bar{x})}{n}$ and $\frac{\sum|x-\bar{x}|}{n}$ for the rainfall data for Pretoria and Durban

	Pretoria			Durban		
Month	x	$(x-\bar{x})$	$\lvert x-\bar{x}\rvert$	x	$(x-\bar{x})$	$\lvert x-\bar{x}\rvert$
Nov	119.6	3.38	3.38	110.9	-0.7	0.7
Dec	119.9	3.68	3.68	106.6	-5.0	5.0
Jan	134.4	18.18	18.18	110.1	-1.5	1.5
Feb	113.0	–3.22	3.22	124.7	13.1	13.1
March	94.2	–22.02	22.02	105.7	-5.9	5.9
		0.00	50.48		0.0	26.20

Pretoria: $\bar{x}$ = 116.22 ma = 10.10 Durban: $\bar{x}$ = 111.60 ma = 5.24

Although the mean deviation gives an accurate figure for the spread of the data set, it is still not very popular as it is difficult to handle mathematically in further statistical analyses. It is also clear that the mean deviation requires considerable calculations if a large ungrouped data set has to be analysed.

A second method of eliminating the negative signs of $(x_1 - \bar{x})$, $(x_2 - \bar{x})$, ..., $(x_n - \bar{x})$ is to use the squares of these deviations and to define the mean square of these, namely

$$m_2 = \frac{\sum(x-\bar{x})^2}{n}$$

as the measure of spread. This measure is known as the second order central sample moment (see Section 5) and is sometimes used as a measure of spread, but the next related measure, known as the standard deviation, s, is more commonly used.

The **standard deviation** of $x_1, x_2, \ldots, x_n$ is defined by

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

There are two important reasons why s is preferred to m_2 as a measure of spread. In the first place s, in contrast to m_2, is indicated in the same units as the x values on which it is based. Thus for example, s for the rainfall data would be given in millimetres, while m_2 will be given in (millimetres)^2. Secondly the sum of squares is divided by $(n-1)$ when s is being calculated, as it renders a more accurate estimate of the population standard deviation. The latter aspect is discussed fully in Chapter 11.

The above definition formula for s is not normally used for calculation purposes. The calculation formula for s is given by

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$$

EXAMPLE 5.3

To illustrate the equivalence of the two formulas for s and at the same time the advantage of using the calculation formula, the standard deviations for the two rainfall data sets (see Table 5.1) are calculated. Table 5.4 contains the calculations.

Table 5.4 Calculation of the standard deviations for the rainfall data for Pretoria and Durban

Pretoria			Durban		
x	$(x - \bar{x})^2$	x^2	x	$(x - \bar{x})^2$	x^2
119.6	11.4244	14 304.16	110.9	0.49	12 298.81
119.9	13.5424	14 376.01	106.6	25.00	11 363.56
134.4	330.5124	18 063.36	110.1	2.25	12 122.01
113.0	10.3684	12 769.00	124.7	171.61	15 550.09
94.2	484.8804	8 873.64	105.7	34.81	11 172.49
581.1	850.7280	68 386.17	558.0	234.16	62 506.96

Definition formula:

$$s = \sqrt{\frac{850.7280}{4}}$$

$$= 14.5836$$

$$\approx 14.58$$

Calculation formula:

$$s = \sqrt{\frac{1}{4}\left\{68\,386.17 - \frac{(581.1)^2}{5}\right\}}$$

$$= 14.5836$$

$$\approx 14.58$$

Definition formula:

$$s = \sqrt{\frac{234.16}{4}}$$

$$= 7.6511$$

$$\approx 7.65$$

Calculation formula:

$$s = \sqrt{\frac{1}{4}\left\{62\,506.96 - \frac{(558.0)^2}{5}\right\}}$$

$$= 7.6511$$

$$\approx 7.65$$

From now on only the calculation formula for s will be used. From Table 5.4 it follows that after squaring the original x values or the $(x - \bar{x})$ values, there are in some cases more significant figures after the decimal comma. In order to calculate the standard deviation accurately, these significant figures should be retained in all the steps of the calculation. It is, however, customary to approximate s after the final calculation to one significant figure more than the original x observations.

The fact that x and ma differ for the same data set (for example, for Pretoria $s = 14.58$ and $ma = 10.10$) follows from the fact that the two measures depict the spread in different ways. However we can conclude from both these measures that the figures for Pretoria's rainfall vary more than those for Durban.

EXAMPLE 5.4

For the price per square metre of the 100 residences in the Johannesburg area (see Table 2.1),

$$\Sigma x = 90\,931.60; \quad \Sigma x^2 = 94\,061\,229.3814.$$

From this it follows that

$$s = \sqrt{\frac{94\,061\,229.3814 - \frac{(90\,931.60)^2}{100}}{99}}$$

$$= 338.97753$$

$$\approx 338.98 \text{ rand}$$

Notes

(a) The standard deviation as discussed above is a descriptive measure of spread for a sample of observations. The standard deviation for a population of size N can similarly be defined as

$$\sigma = \frac{\Sigma (x - \mu)^2}{N}$$

$$= \frac{\Sigma x^2}{N} - \mu^2$$

where μ is defined in Chapter 4, Section 2.1, Note (b) as the population mean. Like μ, σ is a constant value for a population, in contrast to s, which varies from one sample of the population to the next.

(b) In mathematical calculations it is often simpler to work with the square of s, namely the **variance** s^2. An important example of this is the following:

Take k samples, for which only the sizes n_i, the arithmetic means $\bar{x}_i$ and the variances s_i^2 are known. The arithmetic mean and the variance of the joint sample of the $n = n_1 + n_2 + \ldots + n_k$ values are respectively represented by

$$\bar{x}_w = \frac{\Sigma n_i \bar{x}_i}{n} \quad \text{(see Chapter 4, Section 2.2)}$$

$$s_w^2 = \frac{\Sigma (n_i - 1) s_i^2 + \Sigma n_i (\bar{x}_i - \bar{x}_w)^2}{n - 1}$$

The application of this measure to data is illustrated in Example 5.5 below.

EXAMPLE 5.5

The arithmetic means and standard deviations for the turnover values for the 14 'special' and the 16 'ordinary' days in April 1991 at the steakhouse in Example 5.2 are given in the table below:

	Special days	Ordinary days
n	14	16
$\bar{x}$	R1 796.67	R894.27
s	R 745.63	R375.94

Application of the above formulas to the combined group of thirty days gives

$$\bar{x}_w = \frac{14 \times 1\,796.67 + 16 \times 894.27}{30}$$
$$= \text{R}1\,315.39$$

$$s_w^2 = \tfrac{1}{29}\{13(745.63)^2 + 15(375.94)^2 + 14(1\,796.67 - 1\,315.39)^2 + 16(894.27 - 1\,315.39)^2\}$$
$$= 531\,992.9489$$

so that

$$s_w = \text{R}\,729.38$$

2.3.1 Properties of the standard deviation

The fact that the standard deviation is based on all the values has both advantages and disadvantages. It is beneficial in the sense that it fluctuates very little from one sample to the next taken from the same population. The disadvantage in using all observations in the calculation of s, lies in the fact that outliers have a greater influence than would be the case with, for example, the quartile deviation.

The standard deviation is more difficult to interpret than the mean deviation and more complicated than the range and quartile deviation. One of the mathematical characteristics of the standard deviation that is not shared by the other measures of spread is the fact that it is possible to determine the joint standard deviation from the different data sets. The formal relation between the joint and individual standard deviations was given in Note (b) above and illustrated in Example 5.5.

Despite the few disadvantages indicated here, the standard deviation remains the most important measure of spread and it will be used frequently in the remaining chapters of this book.

3 Measures of spread: grouped data

3.1 The range and quartile deviation

The **range of a grouped data set** is defined as the difference between the upper and lower class boundaries of the frequency table.

For the price per square metre of the 100 residences as contained in Table 3.7, the range is

$$r = 1\,700 - 100$$
$$= 1\,600 \text{ rand.}$$

It is clear that the range for a grouped data set will always exceed the range for the corresponding raw data set.

The formula for the quartile deviation for grouped data is identical to the formula for ungrouped data, with the difference that q_1 and q_3 are now derived from the cumulative frequency polygon. With the aid of the cumulative frequencies in Table 4.2 the finding for the ages of 10 691 victims of road traffic collisions in 1988 is that $q_1 = 22.91$ and $q_3 = 41.78$, so that the quartile deviation = 9.4 years.

3.2 The standard deviation

As in the calculation of the arithmetic mean for grouped data, the data set used here has n values divided into k classes, with x and f as the respective class marks and class frequencies.

The **standard deviation for grouped** data is defined by

$$s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n - 1}}$$

The calculation formula is given as

$$s = \sqrt{\frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n - 1}}$$

EXAMPLE 5.6

Assume that the standard deviation for the 100 house prices per square metre, as summarized in Table 3.7, has to be calculated. From Table 5.5 it is clear that $\sum fx = 91\,400$ and $\sum fx^2 = 95\,080\,000$, so that

$$s = \sqrt{\frac{95\,080\,000 - \frac{(91\,400)^2}{100}}{99}}$$
$$= 341.42 \text{ rand.}$$

Table 5.5 Calculation table for the standard deviation for grouped data

Class intervals	x	f	fx	fx^2
[100, 300)	200	2	400	80 000
[300, 500)	400	8	3 200	1 280 000
[500, 700)	600	18	10 800	6 480 000
[700, 900)	800	25	20 000	16 000 000
[900, 1 100)	1 000	19	19 000	19 000 000
[1 100, 1 300)	1 200	12	14 400	17 280 000
[1 300, 1 500)	1 400	10	14 000	19 600 000
[1 500, 1 700)	1 600	6	9 600	15 360 000
			91 400	95 080 000

Notes

(a) As in the case of $\bar{x}$, there are several methods for simplifying the calculation of s. One of the methods involves decreasing each class midpoint x by a constant, say a. As in the case of ungrouped data, this does not affect the value of s. This method is particularly useful if all the class midpoints are numerically very large and the differences between them are relatively small. A second method that is essential in a grouped data set with class marks that are numerically very large and widely scattered, is to subtract a constant, a, from each class mark and to divide the differences by c, the common class width. Thus we transform to a new variable, namely

$$y = \frac{(x - a)}{c}$$

If $\bar{x}$ and $\bar{y}$ respectively represent the arithmetic means, and s_x and s_y respectively represent the standard deviations of the x and y values, it follows that

$$\bar{x} = c\bar{y} + a$$

$$s_x = cs_y$$

This transformation technique is illustrated in Example 5.7, just after these notes.

(b) As mentioned in Note (a) above, if all observations are decreased or increased by the same constant this does not affect the value of the standard deviation (a does not appear in the formula $s_x = cs_y$). This argument applies to all the measures of spread that have been discussed.

(c) When calculating s we of necessity assumed that all the observations in a certain class interval are equal to the class midpoint (assumption for grouped data, Chapter 4, Section 3). In order to compensate partially for this error, a correction, known as **Sheppard's correction**, can be made in the value of s, provided that all the class intervals are of equal length c. Formally it can be expressed as follows:

$$s(\text{corrected}) = \sqrt{s^2(\text{uncorrected}) - \frac{c^2}{12}}$$

Applied to the data in Example 5.6, we find that

$$s(\text{corrected}) = \sqrt{(341.42)^2 - \frac{(200)^2}{12}}$$
$$= 336.50 \text{ rand}.$$

This value is more accurate than s (uncorrected) if it is compared to the actual value of $s = 338.98$, which was found for the raw data.

(d) It is obvious that, for a grouped data set with open-ended class intervals, s cannot be calculated (as is also the case for $\bar{x}$).

(e) There are various relations between measures of spread. Only one, namely a general and simple relation between the range and the standard deviation for data sets that are reasonably symmetrical with a single mode need be mentioned here:

$$r \approx 6 \times s$$

EXAMPLE 5.7

The frequency distribution of the birth mass of 440 babies is given in Table 5.6. The mean and standard deviation of the birth mass of babies will be calculated by means of the transformation

$$y = \frac{x - 3.375}{0.25}$$

Table 5.6 Calculation table for mean and standard deviation of the birth mass of 440 babies

Mass	f	$100\frac{f}{n}$	x	y	fy	fy^2
[2.00, 2.25)	3	0.68	2.125	–5	–15	75
[2.25, 2.50)	8	1.82	2.375	–4	–32	128
[2.50, 2.75)	22	5.00	2.625	–3	–66	198
[2.75, 3.00)	46	10.45	2.875	–2	–92	184
[3.00, 3.25)	98	22.27	3.125	–1	–98	98
[3.25, 3.50)	91	20.68	3.375	0	0	0
[3.50, 3.75)	99	22.50	3.625	1	99	99
[3.75, 4.00)	37	8.41	3.875	2	74	148
[4.00, 4.25)	21	4.77	4.125	3	63	189
[4.25, 4.50)	10	2.27	4.375	4	40	160
[4.50, 4.75)	5	1.14	4.625	5	25	125
	440	100			–2	1 404

The mean and standard deviation of the y values are:

$$\bar{y} = \frac{-2}{440}$$
$$= -0.0045455 \text{ and}$$

$$s_y = \sqrt{\frac{1\,404 - \frac{(-2)^2}{440}}{439}}$$
$$= 1.78834$$

The mean birth mass is

$$\bar{x} = c\bar{y} + a$$
$$= 0.25(-0.0045455) + 3.375$$
$$= 3.37386$$
$$\approx 3.38 \text{ kg}$$

The standard deviation of birth masses is

$$s_x = c \times s_y$$
$$= 0.25(1.78834)$$
$$= 0.44708$$
$$\approx 0.45 \text{ kg}$$

It is also interesting to note that $6(0.45) = 2.70$, which is approximately equal to the range of $4.75 - 2.00 = 2.75$ (see Note (e) above).

4 Interpretation of and uses for the standard deviation

4.1 Interpretation of the standard deviation

As indicated above, the standard deviation, s, is more difficult to interpret than, for instance, the arithmetic mean. The standard deviation reflects the distances of all the individual values from the arithmetic mean, and the greater s is, the further, on average, the values lie from the arithmetic mean. It is known that, for reasonably symmetrical bell-shaped distributions, approximately 68% of the values fall between $\bar{x} - s$ and $\bar{x} + s$, while approximately 95% of all values fall between $\bar{x} - 2s$ and $\bar{x} + 2s$, and 99% between $\bar{x} - 3s$ and $\bar{x} + 3s$.

EXAMPLE 5.8

In Figure 5.5 the relative frequency histogram of the birth masses of 440 babies is recorded. As the distribution is relatively symmetrical around the mean of 3.38, the above note on the percentages between $\bar{x} \pm s$, $\bar{x} \pm 2s$ and $\bar{x} \pm 3s$ should be applicable. This similarity is clear from Table 5.7 and Figure 5.5. We note, for example, that 65% of the values in Table 5.6 fall in the interval [3.00, 3.75) (a combination of the three middle intervals), while the interval with boundaries of $\bar{x} \pm s$ corresponds approximately with the former interval and the approximate expected percentage is also the same.

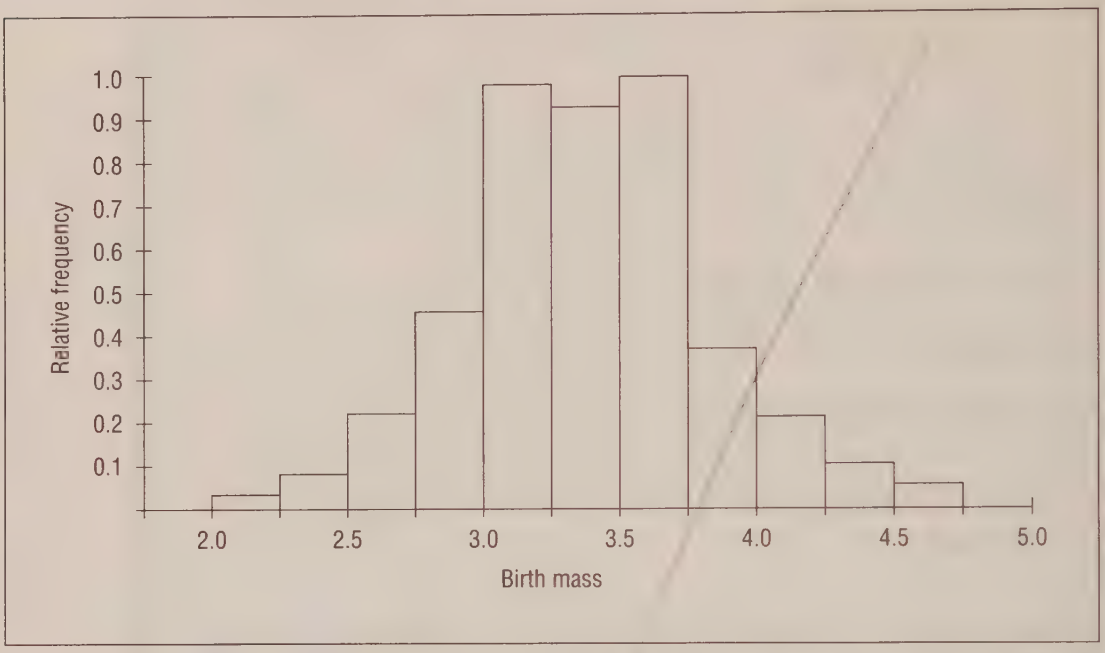


Figure 5.5 Relative frequency histogram of birth masses of 440 babies

Table 5.7 Comparison of observed and expected percentages of the distribution of the birth mass of 440 babies

k	Observed interval (Table 5.6)	Observed percentage	$\bar{x} \pm ks$	Expected percentage
1	[3.00, 3.75)	65	(2.93, 3.83)	68
2	[2.50, 4.25)	94	(2.48, 4.28)	95
3	[2.00, 4.75)	100	(2.03, 4.73)	99

A general rule that applies to any population of values for which the standard deviation is finite, is **Chebyshev**'s theorem that can be formulated as follows:

At least $100\left(1 - \frac{1}{k^2}\right)\%$ of all the values in a population with a mean of μ and a standard deviation of σ will fall between $\mu - k\sigma$ and $\mu + k\sigma$.

EXAMPLE 5.9

The flying squad in a certain city receives an average of ten calls per hour during busy periods and estimates the standard deviation of the number of calls at about 2.5 per hour.

(a) During which part of the busy periods are between six and fourteen calls per hour received?
(b) Determine the boundaries c and d between which the middle 80% of the hourly calls will fall.

(a) With respect to the first question we can argue as follows:

$$\begin{aligned}(\mu + k\sigma) - (\mu - k\sigma) &= 14 - 6\\ &= 8\end{aligned}$$

$$\text{so that } 2(k\sigma) = 8$$

$$\text{or } 2k(2.5) = 8$$

$$\text{Therefore } k = 1.6$$

As a result

$$\begin{aligned}1 - \frac{1}{k^2} &= 1 - \frac{1}{2.56}\\ &= 0.609\\ &\approx 0.61\end{aligned}$$

so that in a busy period for 61% of the time between six and fourteen calls per hour are received.

(b) Here

$$1 - \frac{1}{k^2} = 0.80$$

$$\text{that is } \frac{1}{k^2} = 0.20$$

$$\text{and so } k = \sqrt{5}$$

$$\approx 2.236$$

For 80% of the time the number of calls will therefore lie between

$$10 - 2.236(2.5) \text{ and } 10 + 2.236(2.5);$$

i.e. 4.41 and 15.59 or between 4 and 16 calls per hour.

4.2 Relative measures of spread

Five random samples of size 10 each were drawn from a population of adult males and their heights (in cms) determined. For each sample the standard deviation and the quartile deviation were subsequently determined. The data is given in Table 5.8, together with the means and standard deviations for the five values for each of the two measures.

Table 5.8 Comparison of variation in standard deviations and quartile deviations for five samples of heights of adult males

Sample number	Standard deviation, s	Quartile deviation, qd
1	11.1	8.0
2	8.5	5.0
3	8.9	6.5
4	7.0	3.7
5	10.1	6.3
Mean	9.12	5.9
Standard deviation	1.56589	1.62635

From this table it is clear that the standard deviation for the five s values is lower than that for the 5 qd values, although at first glance the variation of the s values does not seem to be much less than that of the qd values. However, intuition dictates that a comparison of the spread of data sets that differ from each other with regard to order size, would not be justified. The question asked should therefore rather be: How does $s = 1.56589$ for data for which $\bar{x} = 9.12$ compare with $s = 1.62635$ for data for which $\bar{x} = 5.9$? Formally, this comparison can be made by calculating a so-called **relative measure of spread** for both data sets.

The best-known relative measure of spread is the variation coefficient v.

The **variation coefficient** represents the standard deviation as a percentage of the arithmetic mean, i.e.

$$v = 100\left(\frac{s}{\bar{x}}\right)$$

The comparison of the variation in s values with that in qd values can now be made with the aid of the variation coefficient.

$$s \text{ values: } v_1 = 100\left(\frac{1.56589}{9.12}\right)$$
$$= 17.2$$

$$qd \text{ values: } v_2 = 100\left(\frac{1.62635}{5.9}\right)$$
$$= 27.6$$

The relative spread of the s values is therefore considerably less than that of the qd values.

Note

Another less-known relative measure of spread uses the information contained in the quartiles.

The **coefficient of quartile variation** (v_q) is defined by

$$v_q = 100\left(\frac{q_3 - q_1}{q_3 + q_1}\right)$$

4.3 Standardization

Individual values from one data set often have to be compared with individual values from another data set. In psychological testing, for example, the result of some or other test that a person has completed needs to be compared with standard results. This problem is solved by using the technique of standardization. This is a process by means of which the relative size of a value is determined in terms of the mean and standard deviation for comparable data.

Standardization of a number x is accomplished by applying the transformation

$$z = \frac{x - \bar{x}}{s}$$

where z is the so-called **standard value** of x and expresses x in **standard units**.

EXAMPLE 5.10

The South African Weather Bureau publishes meteorological data on weather conditions observed over a protracted period in the various main centres of the country. The monthly minimum temperatures for Bloemfontein and Johannesburg are given in Table 5.9. The means and standard deviations for the minimum temperatures for the two cities are the following:

	Bloemfontein	Johannesburg
$\bar{x}$	8.50833	10.00833
s	5.51419	3.78525

The standard or z values for the minimum temperatures are also given in the table. The z value for Bloemfontein's January minimum is, for example, calculated as follows:

$$z = \frac{15.4 - 8.50833}{5.51419}$$

$$= 1.24981$$

$$\approx 1.25$$

Table 5.9 Monthly minimum temperatures and standard values for Bloemfontein and Johannesburg

Month	Bloemfontein minimum	Bloemfontein z-value	Johannesburg minimum	Johannesburg z-value
Jan	15.4	1.25	14.3	1.13
Feb	14.9	1.16	13.9	1.03
March	12.7	0.76	12.8	0.74
April	8.4	−0.02	10.3	0.08
May	4.0	−0.82	6.9	−0.82
June	0.7	−1.42	4.1	−1.56
July	0.3	−1.49	4.1	−1.56
Aug	3.2	−0.96	6.5	−0.93
Sept	6.3	−0.40	9.1	−0.24
Oct	10.0	0.27	11.8	0.47
Nov	12.0	0.63	12.6	0.68
Dec	14.2	1.03	13.7	0.98
Mean	8.50833	0.0	10.00833	0.0
Std deviation	5.51419	1.0	3.78525	1.0

A number of interesting conclusions that can be drawn from this table are the following:

The minimum temperature of 15.4 degrees recorded in Bloemfontein in January, is 1.25 standard deviations higher than the twelve-month mean of 8.5 degrees.

A monthly minimum of 4.0 degrees in Bloemfontein is comparable to a monthly minimum of 6.9 degrees in Johannesburg (compare the May minimums).

Johannesburg's June and July minimums of 4.1 degrees are relatively lower than the minimums for Bloemfontein for these months, namely 0.7 and 0.3 degrees.

The mean and standard deviation for the z values are 0.0 and 1.0. This is a general property of standard values.

It is important to note that neither relative measures of spread nor standard values are expressed in units. These measures can therefore also be used to draw comparisons between sets of values that are measured in different units. The mean rainfall figure of 84.6 mm for January in Bloemfontein, for example, is 1.29 in standard units ($\bar{x} = 46.7$ and $s = 29.4$), and this figure is approximately equal to the standard value of 1.25 for the minimum temperature of 15.4 degrees in Bloemfontein for January. We can therefore compare the **minimum temperature** for a weather station for a certain month in standard units with the mean **rainfall** for that station and month. Similarly, the variation coefficients for the rainfall figures and monthly minimum temperatures for Bloemfontein can be compared. It follows that for

rainfall figures $$v = 100\left(\frac{29.4}{46.7}\right) = 63.0,$$

minimum temperatures $$v = 100\left(\frac{5.5}{8.5}\right) = 64.7$$

From this we can see that the relative spread in these two variables is similar.

5 Moments

In addition to the characteristics of location and spread, the distribution of observations is also often described in terms of the degree of skewness and peakedness. All four these characteristics are defined in terms of measures that are based on the so-called **sample moments**. The sample moments can be subdivided into **raw moments** and **central moments**.

The rth order **raw moment**, $r = 1, 2, \ldots$, of an **ungrouped data set** $x_1, x_2, \ldots, x_n$ is defined by

$$m'_r = \frac{\sum x^r}{n}$$

and the rth order **central moment**, $r = 1, 2, \ldots$, by

$$m_r = \frac{\sum (x - \bar{x})^r}{n}$$

The rth order **raw moment** of a **grouped data set** with class marks $x_1, x_2, \ldots, x_k$ and class frequencies $f_1, f_2, \ldots, f_k$ (where $n = \sum f$) is defined by

$$m'_r = \frac{\sum f x^r}{n}$$

and the rth order **central moment** of a **grouped data set** by

$$m_r = \frac{\sum f (x - \bar{x})^r}{n}$$

In practice the raw moments are usually calculated first and the central moments are then determined from the following relations:

$$m_1 = 0$$
$$m_2 = m'_2 - (m'_1)^2$$
$$m_3 = m'_3 - 3m'_2 m'_1 + 2(m'_1)^3$$
$$m_4 = m'_4 - 4m'_3 m'_1 + 6m'_2 (m'_1)^2 - 3(m'_1)^4$$

These relations can be proved by using the properties of the summation notation. Note that $m_2 = \frac{n-1}{n} s^2$, where s^2 is the sample variance.

Finally we should like to point out that a transformation of the class marks can often simplify the calculations considerably. This is because the moment expressions contain the class marks up to even the fourth power and these powers still have to be multiplied by frequencies. This transformation has already been discussed in Section 3, Note (a), and is repeated here:

$$y = \frac{x - a}{c}$$

where a = class mark of the interval that falls in the middle (or as close as possible to the middle) of the intervals and c = class width.

The central moments of the y values are calculated first, with the aid of the above moment expressions, and the central moments of the x values are then determined from the following relation, which can also be proved with the aid of the summation notation:

$$m_r(x) = c^r m_r(y)$$

where $m_r(x)$ and $m_r(y)$ are the central moments of the x and y values respectively. The technique is illustrated in the next example (see also Example 5.7).

EXAMPLE 5.11

Consider the frequency distribution of birth age (weeks) of the 440 babies (Appendix A, Data Set 3) mentioned in Example 5.7, as given in Table 5.10. In order to calculate the moments to the fourth power, we shall have to work with numbers like $5(34.5)^4 = 7\,083\,475.31$ and $14(35.5)^4 = 22\,235\,220.086$, which is not only cumbersome, but can also lead to approximation errors.

Table 5.10 Calculation table for the raw moments of the birth ages of 440 babies

x	f	y	fy	fy^2	fy^3	fy^4	$f(1+y)^4$
34.5	5	–5	–25	125	–625	3.125	1 280
35.5	14	–4	–56	224	–896	3.584	1 134
36.5	26	–3	–78	234	–702	2.106	416
37.5	63	–2	–126	252	–504	1.008	63
38.5	107	–1	–107	107	–107	107	0
39.5	100	0	0	0	0	0	100
40.5	81	1	81	81	81	81	1 296
41.5	25	2	50	100	200	400	2 025
42.5	17	3	51	153	459	1.377	4 352
43.5	2	4	8	32	128	512	1 250
	440		–202	1 308	–1 966	12 300	11 916

To simplify calculations, the ages (x values in the first column) were transformed into y values (third column) by means of

$$y = \frac{x - a}{c}$$

where $a = 39.5$ and $c = 1$.

The reader will note that the calculation table above includes a column headed $f(1 + y)^4$. The purpose of this column is to check the accuracy of all the calculations in the table. Using the summation notation we can prove that

$$\Sigma f(1 + y)^4 = \Sigma f + 4\Sigma fy + 6\Sigma fy^2 + 4\Sigma fy^3 + \Sigma fy^4$$

This technique is known as **Charlier's check**.
When applied to this calculation table it therefore follows that:

$$11\,916 = 440 + 4(-202) + 6(1\,308) + 4(-1\,966) + 12\,300$$

The raw moments of the y values are now:

$$m_1' = \frac{\sum fy}{n} = \frac{-202}{440} = -0.45909 \qquad m_3' = \frac{\sum fy^3}{n} = \frac{-1\,966}{440} = -4.46818$$

$$m_2' = \frac{\sum fy^2}{n} = \frac{1\,308}{440} = 2.97273 \qquad m_4' = \frac{\sum fy^4}{n} = \frac{12\,300}{440} = 27.95455$$

By applying the above relations between the central and raw moments it then follows that for the y values:

$$m_2 = 2.97273 - (-0.45909)^2 = 2.76197$$

$$m_3 = -4.46818 - 3(2.97273)(-0.45909) + 2(-0.45909)^3 = -0.56745$$

$$m_4 = 27.95455 - 4(-4.46818)(-0.45909) + 6(2.97273)(-0.45909)^2 - 3(-0.45909)^4 = 23.37536$$

Consequently it follows from the relations $\bar{x} = c\bar{y} + a$ and $m_r(x) = c^r m_r(y)$ that the arithmetic mean and central moments for the x values are represented by:

$$\bar{x} = 1(-0.45909) + 39.5 = 39.04$$

$$m_1(x) = 0$$

$$m_2(x) = (1.0)^2(2.76197) = 2.76197 \approx 2.76, \text{ so that}$$

$$s = \sqrt{\frac{n}{n-1}\, m_2} = 1.66$$

$$m_3(x) = (1.0)^3(-0.56745) = -0.56745 \approx -0.57$$

$$m_4(x) = (1.0)^4(23.37536) = 23.37536 \approx 23.38$$

Note

From the above relations it is clear that the measures of location and spread, $\bar{x}$ and s, are based on the first- and second-order moments. In the next section we shall show that the third- and fourth-order moments can be used to define measures of symmetry and kurtosis respectively.

6 Measures of symmetry and kurtosis

In the current and in previous chapters descriptive measures describing two of the shape characteristics of a data set, namely location and spread were fully explained. In this section we look briefly at descriptive measures of the other two characteristics, namely **symmetry** and **kurtosis**.

6.1 Symmetry

Study the general form of the distribution of a data set as indicated in Figure 5.6. This distribution is skewed, with a long tail end on the positive side.

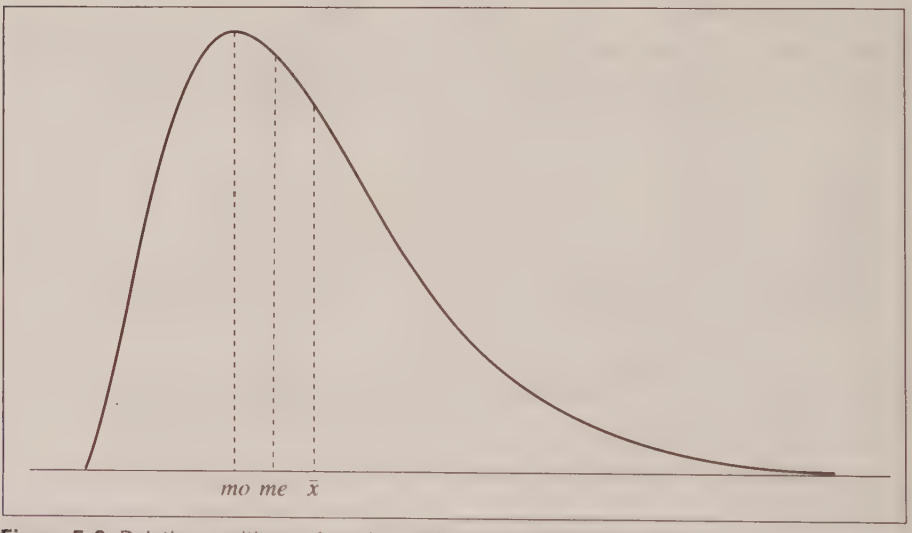


Figure 5.6 Relative positions of mode, median and arithmetic mean

For positively skewed distributions such as this

$$\bar{x} > me > mo$$

For negatively skewed distributions the reverse relation applies:

$$\bar{x} < me < mo$$

and for a fully symmetrical distribution

$$\bar{x} = me = mo$$

It therefore seems to be logical to use the difference between $\bar{x}$ and mo in the definition of a measure of skewness.

The **Pearson coefficient of skewness** is defined by

$$sk(P) = \frac{\bar{x} - mo}{s}$$

Division by the standard deviation ensures that the measure is independent of any unit of measurement and consequently different data sets can be compared with respect to symmetry. For the distribution in Figure 5.6, $sk(P) > 0$, which indicates that the distribution is positively skewed. If $sk(P) < 0$, the distribution is negatively skewed, and in the case of sk(P) = 0, the distribution is symmetrical. The skewness of the distribution increases as $|\bar{x} - mo|$ and consequently $|sk(P)|$ increases.

A second measure of skewness, based on the difference between $(q_3 - me)$ and $(me - q_1)$, was proposed by Bowley.

Bowley's coefficient of skewness is defined by

$$sk(B) = \frac{q_3 + q_1 - 2me}{q_3 - q_1}$$

The third measure of skewness that is often used in practice is based on the moments and is defined by

$$b_1 = \frac{m_3}{m_2^{\frac{3}{2}}}$$

The motivation for using the latter two measures and the interpretation of the various values concerned are analogous to those of $sk(P)$. The advantage of $sk(B)$ is that it is based solely on quartiles so that it can, for example, be calculated for data sets with open-ended class intervals. The skewness measure often has to be based solely on the moments in which case b_1 would be the appropriate choice.

6.2 Kurtosis

In Chapter 4, this term was identified as describing the degree of peakedness of a distribution. Mathematically, kurtosis is related to the absolute size of the gradient of the distribution curve in the vicinity of the mode. The term kurtosis can therefore be used meaningfully only with regard to unimodal distributions. Distributions A, B and C in Figure 5.7 are all symmetrical and may also correspond approximately with respect to location and spread. However, it is clear that the three distributions differ considerably with respect to the characteristic of kurtosis or peakedness.

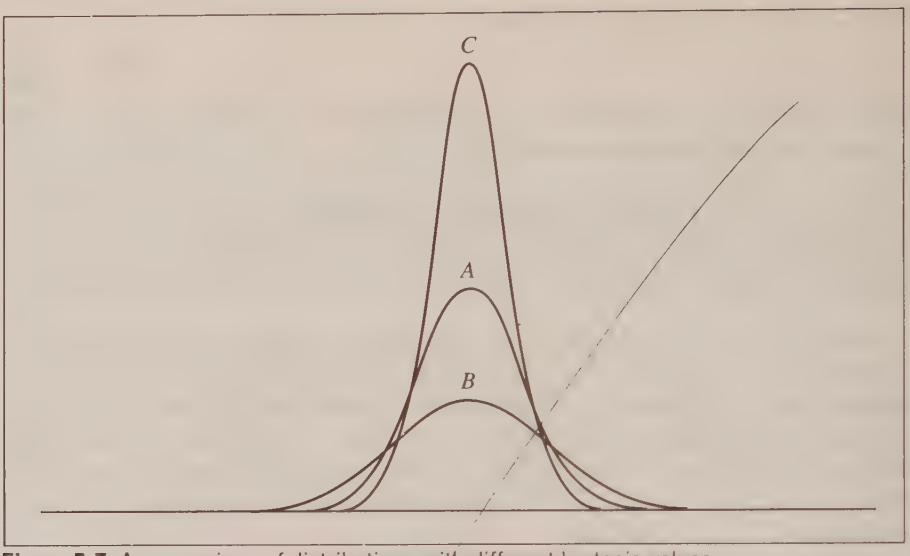


Figure 5.7 A comparison of distributions with different kurtosis values

The customary measure for depicting kurtosis is based on the moments and is known as the coefficient of kurtosis.

The **coefficient of kurtosis** is defined by

$$b_2 = \frac{m_4}{m_2^2} - 3$$

where m_2 and m_4 are the second and fourth central moments respectively of the data.

By means of Figure 5.7 the coefficient of kurtosis can be interpreted as follows:

$b_2 = 0$: normal kurtosis (distribution A)
$b_2 < 0$: negative kurtosis (distribution B)
$b_2 > 0$: positive kurtosis (distribution C).

It is an open question whether any single measure can measure a complex characteristic such as the kurtosis of an observed distribution satisfactorily. Measures that can be more easily calculated than b_2 can therefore be used with as much success to compare distributions mutually with respect to kurtosis. One such measure, based on the quantiles, is given by

$$ku = \frac{q_3 - q_1}{2(d_9 - d_1)} - 0.263$$

It is important to note that, like the coefficients of skewness discussed above, b_2 and *ku* are independent of any unit of measurement.

The calculation of the skewness coefficient b_1 and the kurtosis coefficient b_2 is illustrated in Example 5.12.

EXAMPLE 5.12

Study the frequency distribution of the birth ages of 440 babies as given in Example 5.11. It has already been shown that

$$m_2 = 2.76197,\ m_3 = -0.56745 \text{ and } m_4 = 23.37536$$

so that

$$b_1 = \frac{-0.56745}{(2.76197)^{1.5}}$$
$$= -0.12362$$
$$\approx -0.12$$

$$b_2 = \frac{23.37536}{(2.76197)^2} - 3$$
$$= 0.064223$$
$$\approx 0.06$$

From the above values of b_1 and b_2 it follows that the distribution is virtually symmetrical and has a normal kurtosis.

The moments for calculating measures of symmetry and kurtosis in a data set with open-ended class intervals cannot be determined and we are obliged to use measures based on quantiles. In this regard study Example 5.13.

EXAMPLE 5.13

The ages of South Africans killed in road traffic collisions during 1988 were given in Table 4.2. The reader is encouraged to check the values of the following quantiles:

$d_1 = 12.80$
$q_1 = 22.91$
$me = 31.60$
$q_3 = 41.78$
$d_9 = 53.33$

From this it follows that

$$sk(B) = \frac{41.78 + 22.91 - 2(31.60)}{41.78 - 22.91}$$
$$= 0.790$$

$$ku = \frac{41.78 - 22.91}{2(53.33 - 12.80)} - 0.263$$
$$= -0.03$$

From these two measures it follows that the distribution is somewhat positively skewed with an almost normal kurtosis.

7 Exploratory data analysis

Now that we have discussed a number of descriptive measures, it is fitting that we should again reflect on the first steps that can be followed in handling a raw data set. The basic descriptive methods in the analysis of a data set (as discussed in Chapters 3, 4 and 5) date back to the eighteenth and nineteenth centuries and can be regarded as traditional procedures. Over the past three decades quite a number of new methods have been developed in the field of descriptive statistics. The motivation for much of the research in the field of descriptive statistics is the conviction among scientists that, before any sophisticated analyses can be carried out on a raw data set, the data must first be 'explored', i.e. the data should first be made more accessible by some very simple methods. This approach has led to the fitting description of **exploratory data analysis**. Tukey took the lead in this regard. Tukey's stem-and-leaf display and his box-and-whisker plot are natural, easy and interesting displays that can be used in the exploration of a single-variable data set. They simultaneously reveal the location, spread and symmetry characteristics of a data set.

7.1 The stem-and-leaf display

Table 5.11 reflects the number of beds available, the bed occupancy rate and the financial turnover of the 22 rest camps of the Kruger National Park for the two months of February and July in a certain year. Table 5.12 contains a so-called **stem-and-leaf display** of the data for July. The figures in the second column represent the 'stems' and the figures in the third column the 'leaves'. The first observation, namely 0.74, is written to the right of 0.7* as 4, the second, namely 0.78, to the right of 0.7* as 8, etc. The '*' is a **position marker** and appears in the place of the leaves 0, 1, ..., 9 for a specific stem. Once all of the 22 observations have been represented in this way, we actually have a frequency distribution formed by 0.7*, 0.8*, 0.9* and 1.0* as the classes. The advantages of the stem-and-leaf display are that

(i) it provides a summary of the data;
(ii) the original information is retained; and
(iii) it provides a type of histogram display if the display is turned 90 degrees in an anticlockwise direction.

Table 5.11 Number of beds available, bed occupancy rate and financial turnover of 22 rest camps in the Kruger National Park for two months

Rest camp	Beds available	Occupancy rate		Financial turnover	
		February	July	February	July
Skukuza	18 382	0.56	0.94	591 287	826 958
Pretoriuskop	10 276	0.30	0.90	140 857	387 886
Berg-en-Dal	10 052	0.46	0.92	283 368	449 100
Lower Sabie	6 244	0.71	1.03	289 615	293 054
Satara	12 152	0.66	1.01	415 140	536 709
Olifants Camp	7 451	0.72	0.97	315 894	331 339
Letaba	9 031	0.48	0.91	261 353	391 124
Shingwedzi	8 092	0.20	0.89	66 217	266 730
Punda Maria	1 428	0.43	0.98	44 707	103 414
Crocodile Bridge	1 624	0.35	1.00	28 501	68 307
Orpen	1 232	0.22	0.95	18 653	64 339
Balule	447	0.23	0.88	1 089	2 565
Jock of the Bushveld	336	0.51	0.89	6 336	17 280
Malelane	504	0.13	0.74	3 828	17 250
Nwanedsi	427	0	0.95	0	28 004
Roodewal	532	0	0.79	0	10 640
Boulders	336	0	0.78	0	17 280
Bateleur	952	0.18	0.99	342	26 442
Shimuwini	1 988	0.09	0.92	12 236	57 463
Sirheni	2 210	0.05	0.81	5 277	45 195
Talamati	2 240	0.11	0.86	17 759	51 823
Mbyamiti	1 860	0.10	0.87	9 473	58 684

Data obtained from MBL group OTV00192

Table 5.12 Stem-and-leaf display of the bed occupancy rate of 22 rest camps in the Kruger National Park for July

Stems	Leaves	Frequency
0.7*	4 9 8	3
0.8*	9 8 9 1 6 7	6
0.9*	4 0 2 7 1 8 5 5 9 2	10
1.0*	3 1 0	3
		22

Table 5.13 shows the same stem-and-leaf display, with the difference that the 'leaves' of each 'stem' form an array. This table can now be used to determine any quantile by simply counting off the observation concerned in the desired position. Since the original data was retained, the stem-and-leave array is an array of the raw data set.

Table 5.13 Stem-and-leaf display of the bed occupancy rate of 22 rest camps in the Kruger National Park for July

Stems	Leaves	Frequency
0.7*	4 8 9	3
0.8*	1 6 7 8 9 9	6
0.9*	0 1 2 2 4 5 5 7 8 9	10
1.0*	0 1 3	3
		22

Thus, for example,

$$\begin{aligned} me &= \text{value of the } \frac{22+1}{2} \text{ th observation} \\ &= \text{value of 11.5th observation} \\ &= 0.915 \\ &\approx 0.92 \end{aligned}$$

$$\begin{aligned} q_1 &= \text{value of the } \frac{22+1}{4} \text{ th observation} \\ &= \text{value of 5.75th observation} \\ &= 0.8675 \\ &\approx 0.87 \end{aligned}$$

Similarly we find that $q_3 = 0.97$

Consequently the quartile deviation is

$$\begin{aligned} qd &= \frac{q_3 - q_1}{2} \\ &= \frac{0.97 - 0.87}{2} \\ &= 0.05 \end{aligned}$$

The stem-and-leaf display can be expanded and adjusted in various ways, *inter alia*, to compare two data sets with each other. Example 5.14 illustrates this.

EXAMPLE 5.14

Table 5.13 reflects the number of clients who patronized a steakhouse in Pretoria for April and May 1991 (Appendix A, Data Set 4), together with a classification of the days as 'ordinary' days (indicated by 0) and 'special' days (indicated by 1). See also Example 5.2.

Table 5.14 Type of day and number of clients who visited a steakhouse during April and May 1991

Day	April		May	
	Type of day	Number of clients	Type of day	Number of clients
1	1	112	1	142
2	0	54	0	83
3	0	70	1	112
4	0	65	1	99
5	1	90	1	83
6	1	101	0	35
7	1	80	0	58
8	0	41	1	111
9	0	53	1	87
10	0	68	1	68
11	0	32	1	75
12	1	73	1	134
13	1	82	0	51
14	1	70	0	49
15	0	62	0	43
16	0	48	0	47
17	0	50	1	78
18	0	42	1	80
19	1	63	1	78
20	1	73	0	53
21	1	70	0	54
22	0	42	0	63
23	0	53	0	72
24	0	72	1	92
25	0	89	1	112
26	1	99	1	109
27	1	142	0	83
28	1	159	0	148
29	0	75	1	113
30	1	172	1	155
31			1	138

The number of clients for these two types of days is represented by the back-to-back stem-and-leaf display given in Table 5.15.

Table 5.15 Back-to-back stem-and-leaf display of the number of clients at a steakhouse on 'ordinary' and 'special' days

'Ordinary' days		'Special' days
52	3	
7392281	4	
43183034	5	
3285	6	38
2520	7	30305888
339	8	02370
	9	0992
	10	19
	11	22123
	12	
	13	48
8	14	22
	15	95
	16	
	17	2

The two stem-and-leaf displays can now be compared by making a so-called **five-figure summary** for each type of day:

Type of day	minimum	q_1	*me*	q_3	maximum
Ordinary	32	47.5	54	71	148
Special	63	78	95.5	112.75	172

From this it is clear that special days not only generally have a higher patronage, but also a greater spread in patronage.

By giving the 'stems' a suitable 'thickness', a stem-and-leaf display can further be adjusted with discernment to represent the data set concerned as meaningfully as possible. The following example illustrates this idea of adjustment.

EXAMPLE 5.15

The data below represents the average weekly egg production per hen of a certain egg producer over a period of twenty weeks:

6.0 6.1 6.0 5.9 5.9 5.3 5.5 5.8 4.8 6.0
6.6 6.5 6.6 6.5 6.5 5.8 5.8 5.8 5.7 5.7

Table 5.16(a) shows a stem-and-leaf display of the egg production and it is clear that this display is not suitable, as the stems are too 'thick'. In Table 5.16(b) the problem is solved by creating more stems. In this table '$*$' replaces a 0, 1, 2, 3 or 4,

while '•' replaces a 5, 6, 7, 8 or 9. This display is known as an **extended stem-and-leaf display**.

Table 5.16 Two stem-and-leaf displays of average weekly egg production

(a)		(b)	
4.*	8	4.*	8
5.*	9 9 3 5 8 8 8 8 7 7	5.*	3
6.*	0 1 0 0 6 5 6 5 5	•	9 9 5 8 8 8 8 7 7
		6.*	0 1 0 0
		•	6 5 6 5 5

7.2 Box-and-whisker plot

Figure 5.8 shows a so-called **box-and-whisker plot** of the data in Example 5.14. The rectangle (box) indicates the median as well as the first and third quartiles. The lines and dots ('whiskers') above and below the box cover the rest of the range. A box-and-whisker plot is therefore based on the five-figure summary referred to above. The box-and-whisker plots in this figure clearly indicate the most important differences between the two types of days with respect to location, spread and symmetry:

Ordinary days: Fewer clients, smaller spread, positively skewed
Special days: More clients, larger spread, symmetric

It is also customary to indicate outliers on the box-and-whisker plots. There are several criteria for identifying outliers. Tukey defines an outlier as an observation that is more than one 'jump' outside q_1 or q_3, i.e. an observation larger than $q_3 + t$ or smaller than $q_1 - t$, where

$$t = 1.5(q_3 - q_1)$$
$$= 3qd$$

is the so-called 'jump'. With respect to the data sets being discussed, the following applies to special days:

$$t = 1.5(112.75 - 78)$$
$$= 52.1$$

so that $q_1 - t = 25.9$ and $q_3 + t = 164.9$

Similarly, for ordinary days,

$$t = 35.3,\ q_1 - t = 12.2 \text{ and } q_3 + t = 106.3$$

Consequently in both cases the largest of the observations can be classified as outliers. An outlier is indicated by 'o' in the box-and-whisker plot.

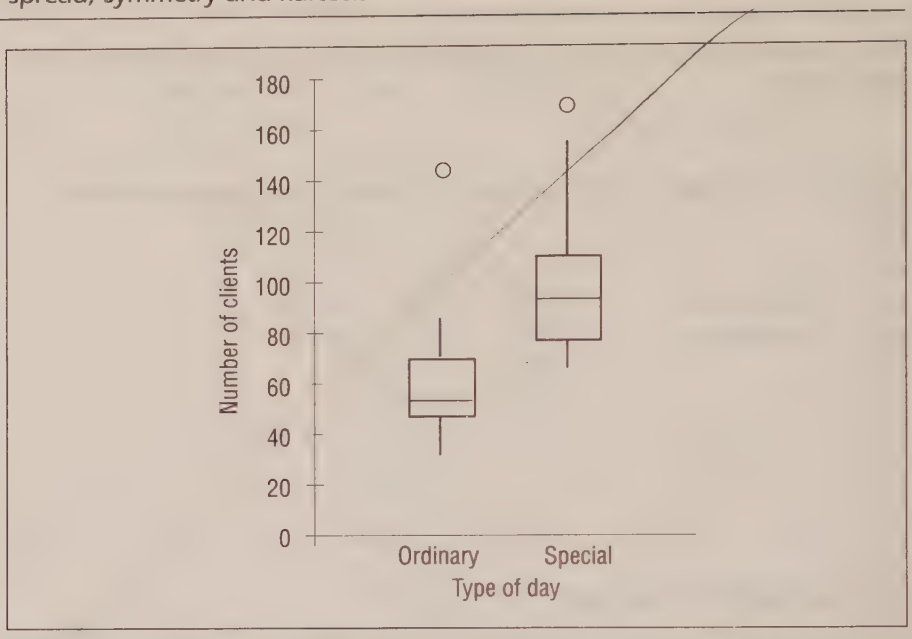


Figure 5.8 Box and whisker plot of the number of clients in a steakhouse on ordinary and special days

Exercises

1 Calculate the range of each of the four samples of selling prices as given in Chapter 4, Exercise 1(a).

2 Determine the quartile deviation, the mean absolute deviation and the standard deviation for the sample of ten selling prices in Durban, as given in Chapter 4, Exercise 1(a).

3 Repeat Exercise 2 for the ten selling prices in Cape Town.

4 Determine the quartile deviation and the standard deviation for the net full supply capacity of the eight regions/catchment areas given in Chapter 4, Exercise 2.

5 Derive the formula for the variance s_w^2 of the combined sample as given in Section 2.3, Note (b), p. 132.

6 Consider a frequency distribution with class marks $x_1, x_2, \ldots, x_k$ and class frequencies $f_1, f_2, \ldots, f_k$ and with $\bar{x}$ and s_x the mean and standard deviation.

Let $y = \frac{x - a}{c}$ with a and c two constant values and x the class marks.

Prove that:

(a) $\bar{x} = c\bar{y} + a$

(b) $s_x = cs_y$

where $\bar{y}$ and s_y are the mean and standard deviation of the y values. (See Section 3.2, Note 3(a), p. 135.)

7 Consider the data in Table 5.11 regarding the five largest rest camps in the Kruger National Park, namely Skukuza, Pretoriuskop, Berg-en-Dal, Satara and Letaba.
   (a) Determine the mean and standard deviation for the bed occupancy rates of each of the five rest camps over the two months.
   (b) Determine from (a) the weighted mean $\bar{x}_w$ and the standard deviation s_w for the five rest camps (see Exercise 5 above). Use the number of beds available as weights.

8 Consider the data on student numbers and achievements as reflected in Chapter 4, Exercise 6. Determine
   (a) the quartile deviation; and
   (b) the standard deviation
   of the student numbers.

9 With reference to the data on the Kruger National Park given in Table 5.11, determine
   (a) the range of the turnover values for each of the two months;
   (b) the mean absolute deviation for the July turnover values of the five largest rest camps (see Exercise 7 above);
   (c) the standard deviation for the July turnover values of these five rest camps; and
   (d) the quartile deviation for the July turnover values of the five rest camps.

10 Consider the frequency distribution of the chrome quantities on 30 steel plates as given in Chapter 3, Table 3.3. Calculate
   (a) the range;
   (b) the quartile deviation; and
   (c) the standard deviation.
   (See also Chapter 4, Exercise 18.)

11 Chapter 3, Table 3.18 reflects a two-way frequency distribution of the ages of workers in the different mining sectors. Calculate, for the ages of workers in the coal mining sector:
   (a) the standard deviation; and
   (b) the quartile deviation.

12 Use the frequency distribution of the taxable income of a certain section of the South African population given in Chapter 4, Exercise 22 and determine the quartile deviation. Can the range or the standard deviation of this data be determined? Give reasons for your answer.

13 In Chapter 4, Exercise 23, the income distribution of three occupational groups is quoted from the 1991 census. Determine the quartile deviation for

(a) the engineering group;
(b) the natural sciences group; and
(c) the medical group.
Which of these groups has the largest variation? Can you explain this?

14 Calculate the coefficient of quartile variation for the income distribution of
(a) the engineering group;
(b) the natural sciences group; and
(c) the medical group
referred to in Exercise 13.

15 If, as in Chapter 4, Exercise 23, we assume that people in the highest income category (R300 000+) earn, on average, R350 000 per annum, calculate:
(a) the standard deviations of the incomes of each of the three occupational groups; and
(b) the standard values for individuals with an annual income of R200 000 in each of the three groups.

16 Determine the variation coefficients for bed occupancy rates over the two months for each of the five largest rest camps in the Kruger National Park referred to in Exercise 7.

17 Consider a sample of size $n = 2$ with values x_1 and x_2. Prove that the variation coefficient is given by

$$v = \frac{\sqrt{2}|x_1 - x_2|}{x_1 + x_2} \times 100$$

18 Compare the standardized ages of 30 year old employees in the three mining sectors by using the data in Chapter 3, Table 3.18 (see also Exercise 11 above).

19 Claim amounts against a specific short-term insurance policy over a single year show a strongly skewed distribution. It is common knowledge that the mean and standard deviation in claim amounts are respectively about R3 000 and R1 200. Determine the percentage of claims that differ from the mean by not more than
(a) R1 800; and
(b) R2 100.
(Clue: Use Chebyshev's inequality.)

20 The treasurer of a sports club decides to round off, to the nearest rand, all the transactions recorded over a period of three months. He is of the opinion that the error, E, which will result, will be positive in some cases and negative in others and that they will cancel each other in the total. It can actually be proved that the total error in the 100 transaction is 0 and that the variance is $\frac{100}{12}$. What percentage of times will the total error

(a) exceed R5; and
(b) exceed R10?
(Clue: Use Chebyshev's theorem.)

21 If the treasurer referred to in Exercise 20 needs to limit the total error of a specific size to 10% of the cases, what will the extent of the error be if he rounds off 50 transactions to the nearest rand?
(Clue: The variance of the total error is now $\frac{50}{12}$.)

22 Calculate, for the frequency distribution of chrome quantities on steel plates as cited in Chapter 3, Table 3.3, the moments m_2, m_3 and m_4 by using the transformation

(a) $y = \frac{x - 68}{4}$

(b) $y = \frac{x - 64}{4}$

23 Consider the age distribution of employees in the different mining sectors referred to in Chapter 3, Table 3.18. Use the transformation $y = \frac{x - 30}{5}$ to determine the first four central moments of the ages of
(a) workers in gold mines; and
(b) workers in diamond mines.

24 Let $x_1, x_2, \ldots, x_k$ and $f_1, f_2, \ldots, f_k$ be the respective class marks and frequencies of a frequency distribution of n values. Let $y = \frac{x - a}{c}$. Prove the following relations:
(a) $\bar{x} = c\bar{y} + a$
(b) $m_r(x) = c^r m_r(y)$

25 Calculate the following measures of symmetry and kurtosis for the frequency distribution of chrome quantities on steel plates in Chapter 3, Table 3.3:
(a) $sk(P)$ and $sk(B)$ (see Chapter 4, Exercise 18 and Chapter 5, Exercise 10).
(b) b_1 and b_2 (see Chapter 5, Exercise 22).

26 With regard to the age distribution of employees in the mining sector as reflected in Chapter 3, Table 3.18, calculate the coefficients of symmetry, b_1 and of kurtosis, b_2, for
(a) workers in gold mines; and
(b) workers in diamond mines (Exercise 23 above).

27 The table below reflects the intelligence quotients of 50 boys:

103	120	124	109	103	107	107	115	103	114
124	117	120	105	110	110	96	137	121	105
98	123	91	97	111	127	132	115	132	97
94	114	115	109	110	98	120	110	119	122
99	128	117	127	116	102	103	117	144	127

(a) Construct a stem-and-leaf display of this data.
(b) Develop the stem-and-leaf array and derive from this me, q_1, q_3 and d_9.
(c) Display the data in the form of a box-and-whisker plot (where there are outliers, they must be indicated).

28 Consider the floor areas of the first 100 houses in Pretoria, as reflected in Appendix A, Data Set 1.
(a) Construct a stem-and-leaf display.
(b) Develop the stem-and-leaf array from (a).
(c) Draw up the five-point summary of the data.
(d) Draw the box-and-whisker plot of the data.

29 Repeat Exercise 28 for the first 100 houses in Cape Town.

30 Develop a back-to-back stem-and-leaf display of the selling prices of the first 25 houses in Durban and Johannesburg respectively as given in Appendix A, Data Set 1.

31 A survey in a company revealed the following age distribution among employees:

Age	Frequency
15 – 19	0
19 – 24	16
25 – 34	37
35 – 44	54
45 – 54	53
55 – 64	11
65 +	0

(a) Calculate the average age for the sample.
(b) Calculate the standard deviation in age.

32 Years of service in a sample of workers are as follows:

Years of service	Frequency
Less than 1	4
1– 3	25
4 – 10	36
11 – 20	61
21 – 30	33
31 – 40	12

(a) Calculate the average length of service in the sample.
(b) Calculate the standard deviation in the period of service.

33 The monthly income of the employees in a company is distributed as follows:

Monthly income	Frequency
R1 000 – R1 499	23
R1 500 – R1 999	33
R2 000 – R2 499	22
R2 500 – R2 999	11
R3 000 – R3 499	13
R3 500 – R3 999	8
R4 000 – R4 499	15
R4 500 – R4 999	4
R5 000 +	31

(a) Calculate the median monthly income.
(b) Calculate the quartile deviation in monthly income.

34 A specific component is used in the manufacture of robot arms. Data on the number of components ordered by clients has been collected over a three-year period. This data is summarized in the frequency table below.

Number of components	Class mark	Number of orders		
		1989	1990	1991
[100, 200)	150	1	12	13
[200, 300)	250	37	27	15
[300, 400)	350	16	2	1
[400, 500)	450	2	3	1
[500, 600)	550	3	8	4
[600, 700)	650	6	3	7
[700, 800)	750	2	3	7
[800, 900)	850	5	5	5
[900, 1 000)	950	3	3	3
[1 000, 1 100)	1 050	2	1	3
[1 100, 1 200)	1 150	2	2	2
[1 200, 1 300)	1 250	1	2	0
[1 300, 1 400)	1 350	0	1	0

Data obtained from MBL group JHB01192

(a) Calculate the arithmetic mean and standard deviation of the number of components per order for the three years.
(b) What conclusions can be drawn from the calculated measures?

35 The amount of beer produced by a brewery for a two-year period is given in Appendix A, Data Set 5.
(a) Calculate the following for the two fiscal years:
  (i) the standard deviation
  (ii) q_1 and q_3
(b) Construct stem-and-leaf displays for the two years under consideration.
(c) What conclusions can be drawn from the measures calculated?

36 The brewery's electricity consumption is given in Appendix A, Data Set 5. Repeat exercise 35 for this variable.

37 The electricity consumption per hectolitre of beer produced (kWh/hl) is given in Appendix A, Data Set 5. Repeat Exercise 35 for this variable.

38 A mining company was approached to investigate the feasibility of lead and zinc mines in an African country. The company is interested in the means and standard deviations of the lead and zinc samples obtained through exploration with a view to investigating the economic feasibility of mining operations in the country. Calculate the arithmetic means and standard deviations of the lead and zinc content of the samples.

Sample	Lead content (kg/ton)	Zinc content (kg/ton)
1	20	60
2	60	40
3	130	90
4	70	120
5	180	70
6	150	110
7	50	150
8	60	30
9	120	60
10	85	70
11	40	70
12	130	100

Data obtained from MBL group RBG00192

39 The branch manager of a supermarket in Zimbabwe investigates the factors that contribute to non-payment of credit, which, in turn, gives rise to cash flow problems in the supermarket. Out of a total of 410 000 accounts at the branch, 10 640 accounts are in arrears. Of these accounts in arrears, 82 accounts have been selected at random.

Income of 82 account holders with accounts in arrears

Income (Z$)	Frequency	Total amount in arrears (Z$)
[300, 520)	10	3 450
[520, 740)	17	6 280
[740, 960)	14	4 930
[960, 1 180)	9	3 580
[1 180, 1 400)	5	1 470
[1 400, 1 620)	9	5 950
[1 620, 1 840)	3	1 620
[1 840, 2 060)	5	3 210
[2 060, 2 280)	4	3 810
[2 280, 2 500)	1	370
[2 500, 2 720)	0	0
[2 720, 2 940)	0	0
[2 940, 3 160)	0	0
[3 160, 3 380)	2	1 220
[3 380, 3 600)	3	3 070
	82	Z$38 950

Data obtained from MBL group HAR00393

(a) Calculate the arithmetic mean and standard deviation of income for the sample of account holders.
(b) Calculate the arithmetic mean of the amount in arrears per account holder as well as the standard deviation of the amount in arrears.
(Note that the amount in arrears per account holder is derived at by dividing the total amount in arrears in each class by the number of account holders in that income group.)

40 During the maintenance of the engines of nineteen DC9 aeroplanes the maximum reverse thrust of the engines was measured. The difference between the port and starboard engines is reflected below:

Calculate the arithmetic mean, standard deviation, minimum, maximum and range of these differences. Also calculate the quartiles q_1, q_2 and q_3.

–120	–115	–100	–90	–45	–40	–20
–20	10	15	30	40	45	45
45	100	100	110	120		

41 Use respondents 501 to 543 in Appendix A, Data Set 2, to compile back-to-back stem-and-leaf displays for the variables Health1 and Health2. What conclusions can be drawn from these displays?

42 Repeat Exercise 41 for the variables Conven1 and Conven2.

43 Repeat Exercise 41 for the variables Financ1 and Financ2.

44 Use respondents 51 to 100 in Appendix A, Data Set 2, to compile box-and-whisker plots for the variables Health1 and Health2. What conclusions can be drawn from these?

45 Repeat Exercise 44 for the variables Conven1 and Conven2.

46 Repeat Exercise 44 for the variables Financ1 and Financ2.

47 Repeat Exercise 44 for the variables Diff1, Diff2 and Diff3.

CHAPTER 6

Fitting curves

1 Introduction

In the three preceding chapters the graphic representation and the calculation of descriptive measures for **univariate data sets** were discussed. In other words, the analysis was limited to a given sample $x_1, x_2, \ldots, x_n$, which was taken from the population being studied. These measurements or observations form a univariate data set and are obtained if the statistical investigation is limited to the analysis of a single population characteristic.

Take, for instance, the population of 843 sales transactions of residences in Johannesburg that were handled by a bank group in 1991 (see Appendix A, Data Set 1). Table 2.1 reflects 100 selling prices that were determined with the aid of a simple random sample from this population. The characteristic concerned here is the selling price, x, at which a specific residence changed hands. However, the population mentioned has a variety of other characteristics that can be determined by means of observation or measurement. Thus the valuation of the buildings, the floor area of the buildings and the size of the stand on which the house is built could all, for example, be important (see columns 4, 5 and 6 in Appendix A, Data Set 1).

To obtain information on two characteristics or variables, namely x and y, a sample consisting of n **pairs** of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ can be drawn. If x and y respectively indicate the selling price (R1 000) and the floor area of the main building (m^2) of the residences, the two-dimensional variable (x_i, y_i) is the information obtained for residence i. This sample is called a **bivariate data set**. Naturally trivariate, four-variate, etc. data sets can be obtained if observations are made on more than two population characteristics. For instance, an analysis can be made of the trivariate sample $(x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_n, y_n, z_n)$, with x, y and z respectively representing the selling price, the valuation of the buildings and the floor areas of the buildings.

Data sets with two and more variables are known as **multivariate data sets**.

In practical situations there is often some or other relation between the observed variables. Wherever possible, researchers try to express these relations between variables in terms of exact mathematical equations. The mathematical relation $y = \frac{V}{x}$, for example, applies between resistance x (ohm) and current y (ampere) at a constant voltage of V volts.

A relation can also be sensed between the valuation of a specific residential unit and the loan amount granted for purchasing that unit. If the valuation is above average, we expect an above-average selling price.

EXAMPLE 6.1

Consider the ten pairs of observations of valuations (x) and selling prices (y) with respect to residences 521 to 530 in Johannesburg (Appendix A, Data Set 1, columns 4 and 3 respectively). For the sake of convenience this bivariate data set is given in Table 6.1. A scatter diagram of the data is given in Figure 6.1. From this figure it is clear that there is an approximately linear relation between the variables y and x in the form $y = a + bx$. The values of a and b, namely the intercept on the y-axis and the slope of the straight line, respectively, can be estimated as follows: With a ruler draw a line in such a way that the line passes all the points as closely as possible. As you know, the equation for a straight line can be determined by means of the co-ordinates of any two points on the line. Assume that the straight line (the dotted line in Figure 6.1) passes through the points A and B with co-ordinates (100, 100) and (300, 320) respectively. The two linear equations from which a and b can be solved are

$$a + 100b = 100$$
$$a + 300b = 320$$

Table 6.1 Valuations (R1 000) and selling prices (R1 000) of 10 residences in Johannesburg

	Valuation (x)	Selling price (y)
1	154.0	160.0
2	285.0	300.0
3	178.0	186.0
4	210.0	230.0
5	109.5	95.0
6	108.0	131.0
7	98.0	107.0
8	75.0	74.4
9	69.5	80.0
10	45.5	47.5

The solution to this equation gives $a = -10$ and $b = 1.1$. If the equation of the estimated straight line is indicated by y, it follows that

$$\hat{y} = -10 + 1.1x$$

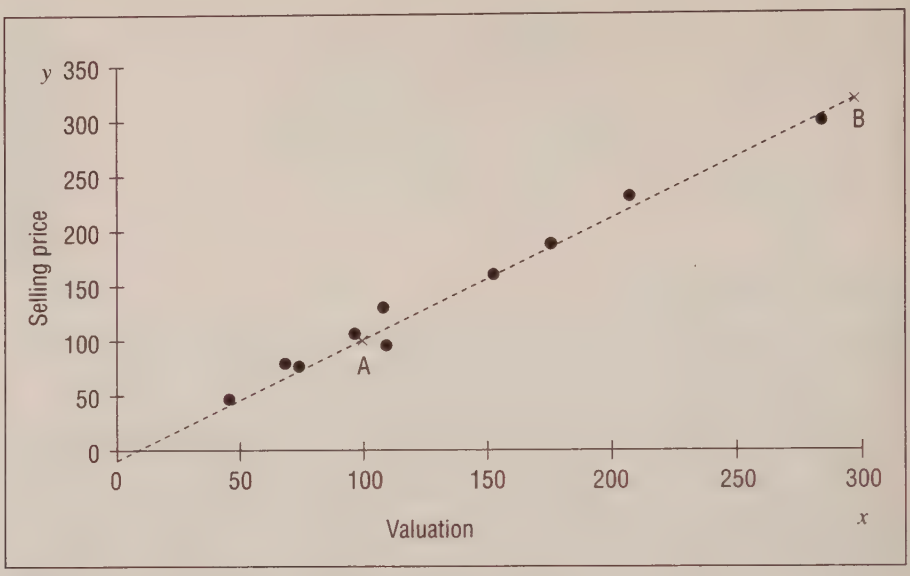


Figure 6.1 Scatter diagram for the data given in Table 6.1

At this stage it should be emphasized that this method, namely **the method of freehand sketching**, gives only an approximation of the linear relation that may exist between two variables x and y. Unless all the observed points lie exactly in a straight line, the freehand sketching method is subjective and can therefore not be used in scientific investigations. The placement of the line on the scatter diagram depends on the judgement of each individual and the equation of the straight line found in this way will therefore vary considerably from one person to the next.

The **linear equation** $y = a + bx$ forms a useful and simple mathematical **model** that is often used to describe the relation between two variables x and y, although the actual relation could be much more complicated, and y could also depend on other factors. In practice, however, the relation is often poorly described by means of a linear equation. The following example illustrates such a situation.

EXAMPLE 6.2

Eight individuals work for an auditing firm. Tape recordings are made during staff meetings to determine how long each staff member speaks during the meetings. The eight employees are now arranged according to seniority from 1 to 8, in such a way that rank 8 is allocated to the head of staff. If the seniority and the percentage of time used by each speaker are indicated by means of x and y respectively, the observations are as indicated in Table 6.2.

Table 6.2 Communication pattern for eight persons

Seniority (x)	Percentage of time spoken (y)
1	0.5
2	3.1
3	2.9
4	7.0
5	9.5
6	17.0
7	22.5
8	37.5

Figure 6.2 is the scatter diagram for this data. From this diagram it is clear that a straight line will give a poor fit for these points. Assume that we decide to fit a parabola with equation $y = a + bx + cx^2$ to the data. As in the case of the linear equation, a freehand sketch of this function can be made on the scatter diagram. By reading three co-ordinate pairs from this sketched function, values can be determined for a, b and c. However, it is obvious that the equation obtained in this way will vary even more from one person to the next than is the case in Example 6.1.

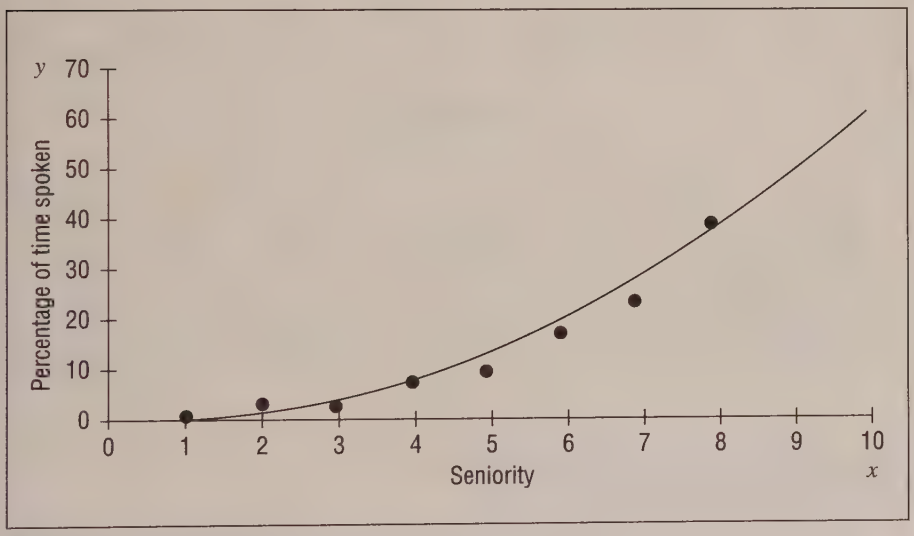


Figure 6.2 Scatter diagram of the data given in Table 6.2

From the above discussion it is clear that we have to decide on a method of curve fitting that is not subject to personal preference and that will simultaneously possess certain better characteristics. The procedure that is usually used for curve fitting is known as the **least squares method**. This method was

developed by the French mathematician A. Legendre at the beginning of the nineteenth century. If a function $f(x)$ provides a good fit to a set of observed points, the least squares method is defined by means of the following:

> The **least squares curve** $\hat{y} = f(x)$ is obtained by minimizing the sum of squares of the vertical distances between the observed points and the corresponding points on the curve.

Figure 6.3 illustrates the principle of least squares: Assume that a straight line has to be fitted to four points. The intercept and slope of the least squares line $\hat{y} = a + bx$ are the values of a and b where

$$\Sigma(y - \hat{y})^2 = \Sigma\{y - (a + bx)\}^2$$

is a minimum.

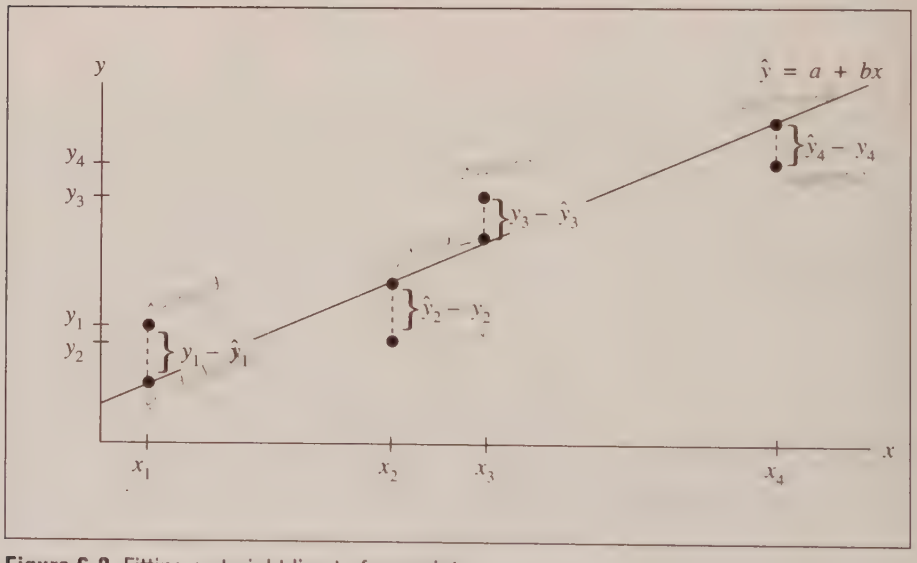


Figure 6.3 Fitting a straight line to four points

2 Why fit a curve?

This question must be answered before we discuss fitting curves to observed data sets in more detail. The following reasons can be given:

2.1 Empirical determination of mathematical laws

Many natural phenomena can be described by some or other mathematical formula that has been determined **empirically** (from practical application) by means of the method of curve fitting. The relation between resistance (x) and current (y) at a constant voltage can, for example, be determined by obtaining a set of points (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) under experimental circumstances and fitting a curve to them.

2.2 Extrapolation

One of the most important objectives of a statistical investigation is to make forecasts on the basis of collected data. A businessman could, for example, be interested in the future value of a piece of land, or the Maize Board in a forecast of the total maize production for a specific season.

If a function $\hat{y} = f(x)$ is a good description of the relation between the variables x and y, it is possible to forecast a y value for a given x value outside the interval of observed x values, provided this relation can still be expected to apply to this x value. This process is known as **extrapolation**.

Assume, for example, that we have fitted a curve to n pairs of (x, y) observations, in which x indicates the speed (km/h) and y the corresponding fuel consumption (ℓ/100 km) of a certain model car. Assume that the fuel consumption was determined at the following x values:

40 60 80 100 120 (km/h)

In this case $\hat{y} = f(140)$ is an estimate of the fuel consumption at a speed of 140 km/h.

It is customary to indicate the variable for which we want to establish a value by means of extrapolation with the symbol y. This variable is called the **dependent variable**, while x is the **independent variable**. Sometimes the terms **explanatory variable** or **predictor** are used instead of **independent variable**.

2.3 Interpolation

It often becomes necessary to estimate a y value for an x value within the interval of observed x values. This process is known as **interpolation**.

Assume that the demand (y) for a certain consumer item is known if the price (x) is 70c, 82c, 90c and 125c respectively. If a curve $\hat{y} = f(x)$ is fitted to the observed data, $\hat{y} = f(115)$ is the estimated demand for the item if the price is 115c. Note that 115c is a value in the interval 70c to 125c and that we have therefore interpolated.

Notes

(a) If a straight line gives a good description of the relation between x and y, as in the case of the data in Example 6.1, this curve can be used confidently for interpolation purposes. The curve can also be used for extrapolation purposes, provided it can be assumed that the same functional relation applies outside the interval of observed x values. In general, however, extrapolation should be undertaken circumspectly since the relation could apply only within the interval of observed x values.

(b) The nature of the data in Example 6.2 is such that the curve that was fitted cannot be used for extrapolation purposes. As the discrete variable x can assume only integer values from 1 to 8, it would not be meaningful to interpolate between these values. However, we can use the curve that was fitted to the data to postulate a mathematical law for this group phenomenon.

3 The linear relation

Assume that a bivariate data set (x_i, y_i), with $i = 1, 2, \ldots, n$ is available and that the scatter diagram on it indicates that there is a linear relation of the form $\hat{y} = a + bx$ between x and y. The least squares values of the intercept and slope are respectively the values of a and b for which

$$\Sigma(y - \hat{y})^2$$

is a minimum. In this regard, review Figure 6.3, in which the case $n = 4$ was discussed. By means of the mathematical technique of differentiation we can indicate that the least squares values of a and b are obtained by solving the following system of equations, known as **normal equations**:

$$\Sigma y = na + b\Sigma x$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

If the first equation is divided by n, it follows after a rearrangement of the terms that

$$a = \bar{y} - b\bar{x}$$

Subsequently, if we multiply the first equation by $\frac{\Sigma x}{n}$ and subtract the resulting equation from the second equation it follows that

$$b = \frac{\Sigma xy - \frac{1}{n}(\Sigma x \Sigma y)}{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}$$

We can therefore use the two equations to fit a straight line to the observed (x, y) points in a scientifically accountable manner.

EXAMPLE 6.3

Consider the fit of a least squares line to the data of Table 6.1. In Table 6.3 $\sum x$, $\sum y$, $\sum xy$ and $\sum x^2$ are calculated as the totals of columns (1) to (4). The calculations in column (5) will only be required in later calculations, while column (6) reflects the estimated values, $\hat{y}$.

Table 6.3 Calculations for the fit of a straight line to the data in Table 6.1

(1) Valuation x	(2) Selling price y	(3) xy	(4) x^2	(5) y^2	(6) $\hat{y}$
154.0	160.0	24 640.00	23 716.00	25 600.00	163.08
285.0	300.0	85 500.00	81 225.00	90 000.00	301.92
178.0	186.0	33 108.00	31 684.00	34 596.00	188.52
210.0	230.0	48 300.00	44 100.00	52 900.00	222.43
109.5	95.0	10 402.50	11 990.25	9 025.00	115.92
108.0	131.0	14 148.00	11 664.00	17 161.00	114.33
98.0	107.0	10 486.00	9 604.00	11 449.00	103.73
75.0	74.4	5 580.00	5 625.00	5 535.36	79.35
69.5	80.0	5 560.00	4 830.25	6 400.00	73.52
45.5	47.5	2 161.25	2 070.25	2 256.25	48.09
1 332.5	1 410.9	239 885.75	226 508.75	254 922.61	1 410.90

$\bar{x} = \frac{1\,332.5}{10} = 133.25; \quad \bar{y} = \frac{1\,410.9}{10} = 141.09$

With the aid of the calculation equation for b, it follows that

$$b = \frac{239\,888.75 - (1\,332.5)(1\,410.9)/10}{226\,508.75 - (1\,332.5)^2/10}$$

$$= \frac{51\,883.325}{48\,953.125}$$

$$= 1.05986$$

Now substitute this value of b in $a = \bar{y} - b\bar{x}$, then it follows that

$$a = 141.09 - 1.05986(133.25)$$
$$= -0.13598$$

The equation of the least squares line is therefore

$$\hat{y} = -0.136 + 1.060x$$

in which both the x and y units are thousands of rands.
It is now easy to calculate the $\hat{y}$ values given in column (6). It thus follows, for instance, that

$$\hat{y}_1 = -0.136 + 1.060(154)$$
$$= 163.08$$

If the $\hat{y}$ values are calculated accurately enough, then

$$\Sigma\hat{y} = \Sigma y$$

This result is generally applicable if a least squares line is fitted to the data.

We could also have solved for a and b from the normal equations

$$1\,410.9 = 10a + 1\,332.5b$$
$$239\,885.75 = 1\,332.5a + 226\,508.75b$$

Assume that we require an estimate of the price at which a house will be sold if the valuation of a residential building is R400 000. This estimate is obtained by substituting $x = 400$ in the equation of the fitted curve:

$$\hat{y} = -0.136 + 1.060(400)$$
$$= 423.864 \text{ thousand rand}$$
$$= \text{R}423\,864$$

Figure 6.4 shows both the least squares and the freehand sketch fits. Note the differences between the a and b values of these two straight lines.

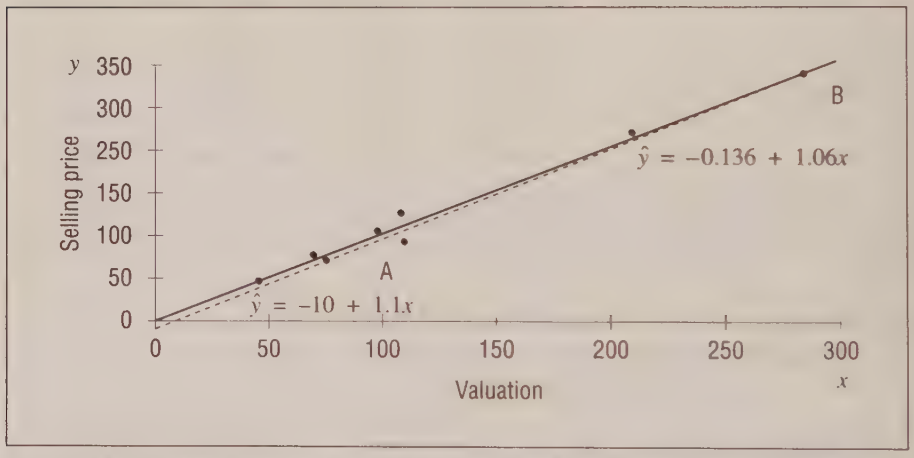


Figure 6.4 Difference between the least squares fit and the freehand-sketch fit for the data in Table 6.1

EXAMPLE 6.4

A spokesman for a catering company maintains that there is a linear relation between the number of institutions that provide accommodation services (x) and the number of institutions that render catering services (y). Table 6.4, columns (1) and (2), reflects the number of institutions that provide accommodation and catering services respectively. This information pertains to the nine development regions of South Africa.

Table 6.4 Number of institutions that provide catering and accommodation services

Region	(1) Accommodation x	(2) Catering y	(3) xy	(4) x^2	(5) y^2
A	570	751	428 070	324 900	564 001
B	128	88	11 264	16 384	7 744
C	175	243	42 525	30 625	59 049
D	416	561	233 376	173 056	314 721
E	448	643	288 064	200 704	413 449
F	98	220	21 560	9 604	48 400
G	64	145	9 280	4 096	21 025
H	256	1 574	402 944	65 536	2 477 476
J	54	131	7 074	2 916	17 161
	2 209	4 356	1 444 157	827 821	3 923 026

$\bar{x} = \frac{2\ 209}{9} = 245.444; \quad \bar{y} = \frac{4\ 356}{9} = 484$

Source: *SA Labour Statistics*, 1991

From the data in Table 6.4 it follows that

$$b = \frac{1\,444\,157 - \frac{1}{9}(2\,209)(4\,356)}{827\,821 - \frac{1}{9}(2\,209)^2}$$

$$= \frac{375\,001}{28\,5634.22}$$

$$= 1.313$$

Therefore

$$a = 484 - 1.313(245.444)$$
$$= 161.732$$

The least squares line is therefore approximated by

$$\hat{y} = 161.73 + 1.31x$$

Notes

(a) When interpreting results, it is important to note that the problem being studied determines whether or not the value of a has any practical meaning. In this regard, consider the results of Example 6.3. If we are prepared to extrapolate outside the interval of observed x values to the value $x = 0$, $\hat{y} = -0.136$, the estimated intercept value. From a mathematical point of view, this means that if a residence is valued at zero rands ($x = 0$), the seller will have to pay the buyer R136 to get rid of his house! As was said before, extrapolation should be undertaken cautiously, as the relation between x and y (if any) does not necessarily remain linear outside the domain of observed x values.

(b) The value of b, the estimated slope, is always of both theoretical and practical importance. Theoretically b, together with a, determines the position of the least squares line. In practice, b measures the mean change of y for a unit change in x. In Example 6.3, $b = 1\,060$ means that there is an average increase of R1 060 in the selling price for each increase of R1 000 in the valuation of a residence. The b value of 1.31 in Example 6.4 means that, for an increase of 100 institutions rendering accommodation services, an average increase of 131 institutions rendering catering services can be expected.

(c) For a straight line passing through the origin, $\hat{y} = bx$, with $b = \dfrac{\sum xy}{\sum x^2}$.

4 The coefficients of determination and correlation

If we use a bivariate data set, it is theoretically always possible to fit a least squares line by calculating $\sum x$, $\sum y$, $\sum xy$ and $\sum x^2$ and to determine values for a and b from these calculations. However, the immediate question is whether there is really a linear relation between x and y and if so, how strong this relation is. The degree of strength of this relation is described by the term **correlation**. Generally, linear correlation means that the relation between x and y is linear, and more specifically positive linear correlation means that for an increase in x, there is an increase in y. The scatter diagram gives an indication of how strong the relation between x and y is. The scatter diagram in Figure 6.5, for example, indicates that there is a perfect positive linear relation between x and y.

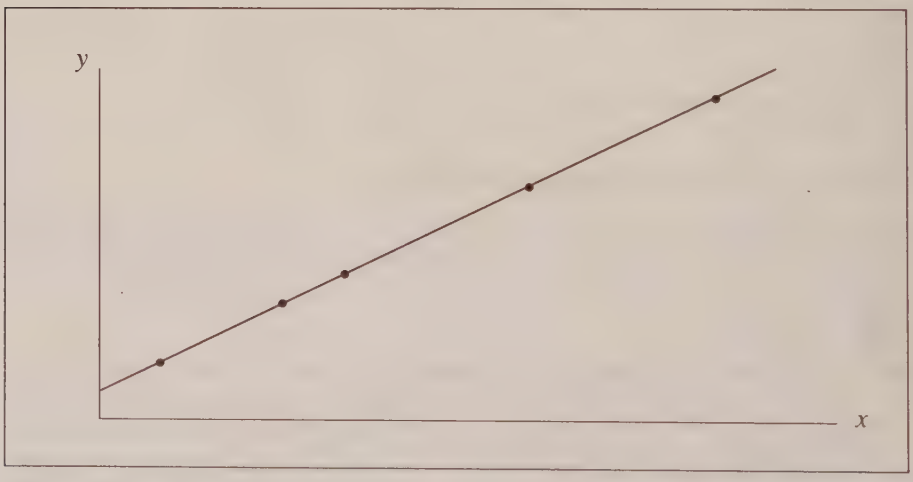


Figure 6.5 Example of data for which there is perfect positive linear correlation between x and y

In Example 6.4 we fitted a straight line without first plotting the (x, y) points. A scatter diagram of this data is given in Figure 6.6.

This figure clearly indicates that the degree of positive linear correlation (or relation) between x and y is relatively low.

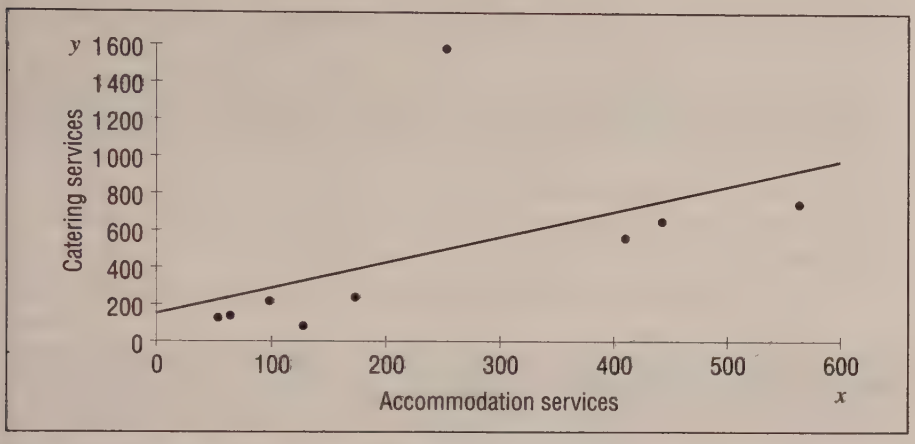


Figure 6.6 Example of data for which there is a poor positive linear correlation between x and y

The establishment of the strength of the relation between two variables is naturally important not only in the linear case, but also in the case of any other type of relation. In view of this, the problem will first be discussed in general terms.

Assume that the least squares curve $\hat{y} = f(x)$ describes the relation between two variables x and y. A suitable measure must be found to quantify the strength of this relation. The following measure is generally used:

The **coefficient of determination** R^2 provides an indication of how well the least squares curve fits the observed data and is defined as

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

If $\hat{y} = f(x)$ provides a perfect fit, $y_i = \hat{y}_i$, $i = 1, 2, \ldots, n$. In such a case $\sum(y - \hat{y})^2 = 0$, and therefore $R^2 = 1$. On the other hand, $\hat{y}_i = \bar{y}$, $i = 1, 2, \ldots, n$ (i.e. $R^2 = 0$), indicates a situation in which there is no relation between x and y. Generally, $R^2 \leq 1$.

4.1 The correlation coefficient r

We shall now show that where $f(x) = a + bx$, the formula for R^2 can be reduced to an expression that can easily be used as a basis for calculation. The square root of this expression is called the correlation coefficient r and is a number between -1 and $+1$.

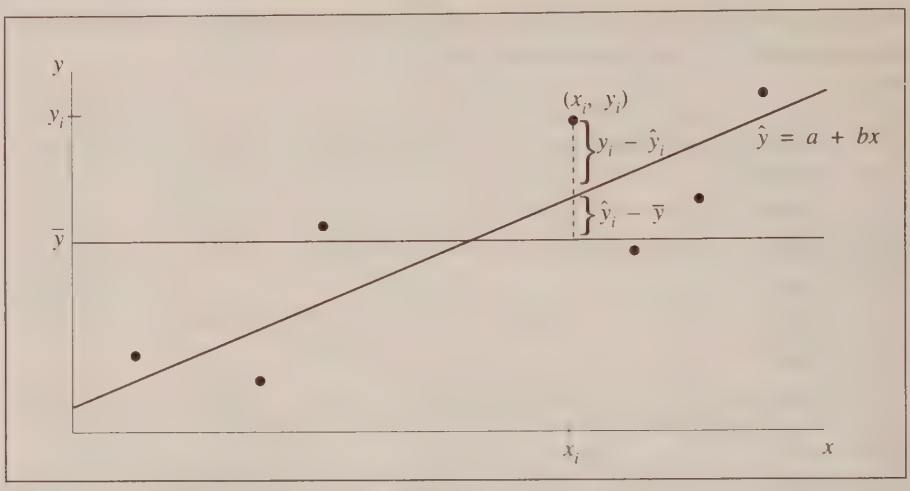


Figure 6.7 Explained and unexplained deviations

Figure 6.7 indicates that for a given value of x the **total deviation** $(y_i - \bar{y})$ can be divided as follows:

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

$(\hat{y}_i - \bar{y})$ is known as the **explained deviation** and $(y_i - \hat{y}_i)$ the **unexplained deviation**. In the case above, namely where $f(x)$ is a linear function of x, it can be shown that the total variation of the y values consists of two components, namely the explained and the unexplained variation:

$$\underset{\text{(total)}}{\Sigma(y - \bar{y})^2} = \underset{\text{(explained)}}{\Sigma(\hat{y} - \bar{y})^2} + \underset{\text{(unexplained)}}{\Sigma(y - \hat{y})^2}$$

In this case the coefficient of determination is therefore expressed as the relation of the explained variation of y to the total variation of y, i.e.

$$R^2 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2}$$

Since $\hat{y}_i = a + bx_i$, where $a = \bar{y} - b\bar{x}$, the explained deviation $(\hat{y}_i - \bar{y}_i)$ can be expressed as

$$\begin{aligned}(\hat{y}_i - \bar{y}_i) &= a + bx_i - \bar{y} \\ &= b(x_i - \bar{x})\end{aligned}$$

The coefficient of determination in the linear case is usually indicated by r^2, and can therefore be expressed as

$$r^2 = \frac{b^2\Sigma(x - \bar{x})^2}{\Sigma(y - \bar{y})^2} = \frac{b^2\left(\Sigma x^2 - \frac{1}{n}(\Sigma x)^2\right)}{\Sigma y^2 - \frac{1}{n}(\Sigma y)^2}$$

As already stated, the square root of this expression is of great practical importance.

The **correlation coefficient** r, which is a measure of linear relationship between two variables x and y, is defined as

$$r = \frac{b\sqrt{\sum x^2 - \frac{1}{n}(\sum x)^2}}{\sqrt{\sum y^2 - \frac{1}{n}(\sum y)^2}}$$

If there is a positive linear relation between x and y, b (the slope) is positive and consequently r will also be positive. In the case of a negative relation, b is negative and r will consequently also be negative. An alternative way of expressing r can be found by dividing the numerator and the denominator by $\sqrt{(n-1)}$ namely

$$r = \frac{bs_x}{s_y}$$

where s_x and s_y respectively represent the standard deviations of the x and y observations.

EXAMPLE 6.5

To illustrate, the correlation coefficient for each of the data sets given in Example 6.3 and 6.4 is calculated.

From the data in Example 6.3 it was found that $b = 1.060$. From the information in Table 6.3 we find that the correlation between valuation (x) and selling price (y) is calculated as follows:

$$r = \frac{1.060\sqrt{48\,953.125}}{\sqrt{254\,922.61 - \frac{1}{10}(1\,410.9)^2}}$$

$$= 0.9923$$

Note that the value of $\sum x^2 - \frac{1}{n}(\sum x)^2$ has been calculated before in the determination of the value of b in Example 6.3. As can be deduced from the scatter diagram in Figure 6.1, the degree of positive correlation between x and y is high.

We now calculate r for the data in Example 6.4. From Table 6.4 and the derived value of b it follows that the correlation between the number of institutions that provide accommodation services (x) and the number of institutions that provide catering services (y) is given by

$$r = \frac{1.313\sqrt{285\,634.19}}{\sqrt{3\,923\,026 - \frac{1}{9}(4\,356)^2}}$$

$$= 0.5209$$

This value for r is considerably smaller than the value for r for Example 6.3. A comparison of the scatter diagram in Figure 6.6 with that in Figure 6.1 also indicates clearly that the linear relation between x and y in the former figure is less obvious than in the latter figure.

From the above example it is clear that an apparently large value for r does not necessarily indicate a strong linear relation between two variables. How large should r then be before we can say that x and y have strong linear correlation? This question can be partially answered by studying the general formula for r^2 again, namely

$$r^2 = \frac{\text{variation in } y \text{ explained by } \hat{y}}{\text{total variation in } y}$$

Mention has already been made of the point that $r^2 = 1$ if the line $\hat{y} = a + bx$ accounts for the total variation of y. The case $r^2 = 0$ corresponds with the situation in which the least squares line cannot account for any variation in y. Thus, for example, it follows for $r^2 = 0.90$ that 90% of the variation in the dependent variable y can be explained by the explanatory (independent) variable x.

In Example 6.5 we calculated the correlation between the number of institutions that provide accommodation services (x) and catering services (y) respectively, and we found a value of $r = 0.52$. From this it follows that $r^2 = 0.27$. Consequently the straight line

$$\hat{y} = 161.73 + 1.31x$$

accounts for only 27% of the variation in the y values. This indicates a relatively poor linear relation between the variables x and y.

In practice, r is often calculated as a first step in the study of the relation between x and y. It is a useful measure for determining whether the fit of a straight line to the data would in fact be meaningful. Since the value of b is unknown in such cases, another calculation formula for r can be found by substituting the least squares value of b, namely

$$b = \frac{\sum xy - \frac{1}{n}\left(\sum x \sum y\right)}{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}$$

in the definition of r. This gives

$$r = \frac{\sum xy - \frac{1}{n}\left(\sum x \sum y\right)}{\sqrt{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}\sqrt{\sum y^2 - \frac{1}{n}\left(\sum y\right)^2}}$$

EXAMPLE 6.6

An experiment was set up to determine the relation between level of sound intensity and distance. The observations in the first two columns of Table 6.5 represent the distance in metres (x) and the level of sound intensity in decibels (y). A 125 cc motorcycle was used as constant sound source and set to idle at 6 000 revolutions per minute. The readings were taken by means of a B & K type 2203 sound level meter that was placed at varying distances from the sound source. In Table 6.5 the totals Σx, Σy, Σxy, Σx^2 and Σy^2 are also calculated. Using this data it follows that

$$r = \frac{60\,158 - \frac{1}{10}(90)(724.5)}{\sqrt{1\,140 - \frac{1}{10}(90)^2}\sqrt{53\,398.13 - \frac{1}{10}(724.5)^2}}$$

$$= \frac{-504.7}{\sqrt{330}\sqrt{908.105}}$$

$$= -0.92 \qquad (r^2 \approx 0.85)$$

Table 6.5 Calculation table for the correlation coefficient between distance (m) and sound intensity level (decibels)

x	y	xy	x^2	y^2
0	95.0	0	0	9 025.00
2	82.0	164.0	4	6 724.00
4	77.0	308.0	16	5 929.00
6	73.5	441.0	36	5 402.25
8	71.0	568.0	64	5 041.00
10	68.5	685.0	100	4 692.25
12	66.3	785.6	144	4 395.69
14	65.0	910.0	196	4 225.00
16	63.7	1 019.2	256	4 057.69
18	62.5	1 125.0	324	3 906.25
90	724.5	6 015.8	1 140	53 398.13

$\bar{x} = \frac{90}{10} = 9.0; \qquad \bar{y} = \frac{724.5}{10} = 72.45$

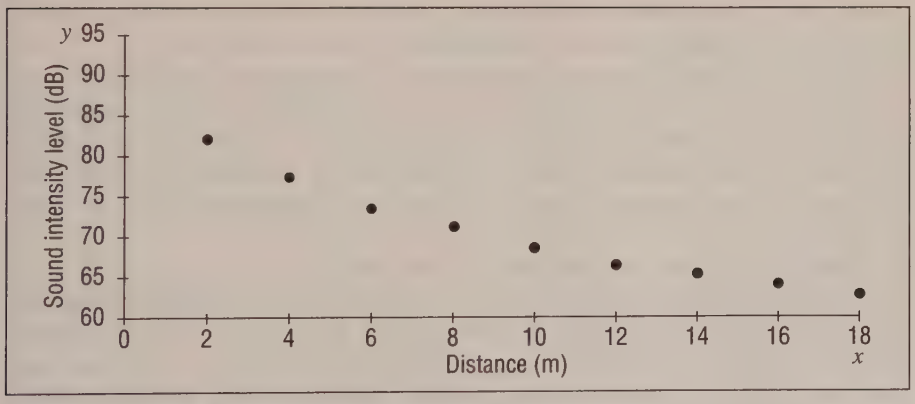


Figure 6.8 Scatter diagram of the data in Table 6.5

Since the value of r is negative, there is a negative linear correlation between x and y. The least squares line $\hat{y} = a + bx$ will therefore explain 85% of the variation in the y values. The scatter diagram of the data is given in Figure 6.8. From this figure it is clear that the observed points lie on an almost smooth curve, but this curve is not a straight line. Although the coefficient of determination is therefore quite high, a suitable non-linear curve should render a much higher value for the coefficient of determination R^2. This aspect is accorded further attention in Section 6.

Notes

(a) It can be shown that scale and locality transformations carried out on the variables x and y do not change the value of the correlation coefficient. Assume that

$$u_i = \frac{x_i - h}{c} \quad \text{and} \quad v_i = \frac{y_i - k}{d}$$

then

$$r_{uv} = r_{xy}$$

The advantage here is that r can be calculated with the aid of smaller order numbers.

(b) Note that the value of r is independent of the units in which x and y are measured. If the distance (x) in Example 6.6 were to be given in centimetres instead of metres, the value of r would not change.

5 Graphic representation of residuals and causal relations

In practice, the strength of the linear relation between the (x, y) observations is often found to be relatively low. The deviations of the y values from the least squares line $\hat{y} = a + bx$ often provides information that can to a large extent explain this state of affairs. A graphic representation of the **residuals** $y - \hat{y}$ against x enables the researcher to study these deviations visually.

Figure 6.9 shows four possible patterns of least squares residuals. It was indicated in the previous section that, if a linear relation is assumed between y and x, $\sum y = \sum \hat{y}$. From this result it follows that $\sum (y - \hat{y}) = 0$. The sum of the residuals, and consequently also the average of the residuals, is therefore 0. This characteristic of the residuals explains why the plotted points fluctuate around a value of 0 on the residual axis (vertical axis).

Figure 6.9(a) shows an idealized residual plot: the residuals appear to be scattered randomly around 0 in a more or less rectangular strip. In contrast, the patterns of residuals in Figures 6.9 (b), (c) and (d) point to phenomena in the data that could cause problems if the fit of a least squares line to this data were maintained. Figure 6.9(b) indicates that the variation of the y values increases with the increase in x. The accuracy of extrapolation of y will therefore decrease progressively as x increases. A non-linear pattern as indicated in Figure 6.9(c) points to a possible non-linear relation between y and x. The aspect of non-linear relations between variables is discussed further in Sections 6 to 8.

In addition to the problems associated with Figures (b) and (c), two additional problems are frequently encountered with the application of the least squares fit, namely those of **outliers** and of **influential observations**. Each of these two concepts will now be discussed briefly.

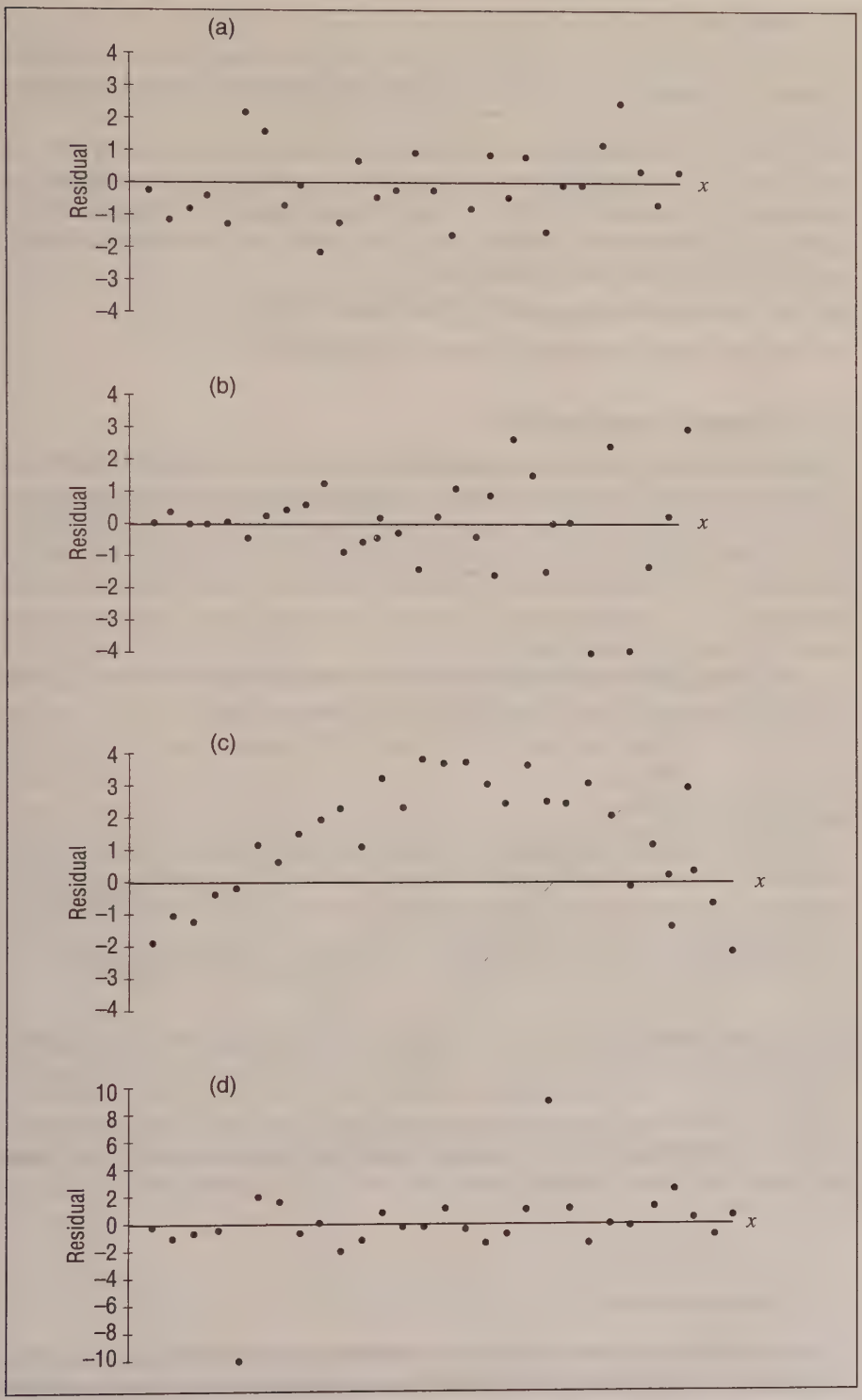


Figure 6.9 Four possible patterns of least squares residuals

5.1 Outliers

Calculation of the a and b values is based on all the observations in the data set. Consequently outliers can exercise considerable influence on the estimated intercept and slope values.

Figure 6.9(d) is a residual plot that indicates the incidence of outliers. In the theory of least squares curve fit an outlier is regarded as a data point that provides a high residual. If the (x, y) pair that corresponds with this data point is removed from the data set and a new least squares curve is fitted to the data, one or both of the following can happen:

(a) The intercept and slope values remain more or less unchanged.

(b) The correlation coefficient increases in absolute value.

5.2 Influential observations

A data point is regarded as influential if its removal from the data set is accompanied by a considerable change in the position of the least squares line, that is, the intercept and slope values change considerably. A shortcoming of least squares curve fitting is that the position of the straight line is strongly influenced by y values that correspond with extreme x values in the data set. This problem can be overcome if it is possible to select the x values proportionately across the value domain of x.

A disadvantage of graphic representation of the residuals against x is that the residual plot cannot always identify influential observations.

EXAMPLE 6.7

A building society employee believes that with residences there is a linear relation between the floor area of the main building (x) and that of the outbuilding (y). The floor areas of the main buildings and outbuildings of ten houses are the following:

x	138	168	500	188	150	166	165	108	162	145
y	0	53	96	35	18	45	16	30	69	42

The values of both x and y are in square metres. For the ten sets of observations the least squares line is $\hat{y} = 5.779 + 0.183x$, while the correlation coefficient is $r = 0.73$.

A study of this table reveals that $x = 500$ can be regarded as an extreme value. The data point (500, 96) can therefore possibly be an influential value. After this point has been removed from the data set, the least squares line is $\hat{y} = -11.688 + 0.297x$, while $r = 0.32$.

5.3 Causal relations

When the relation between variables is examined, the assumption is usually that an increase or decrease in the value of an explanatory variable can cause a change in the value of the dependent variable. It is often difficult, or

impossible, to prove that there is a **causal relation** between certain variables. In practice, two variables are often apparently strongly correlated, despite the fact that there is no causal relation between them. Assume, for example, that strict legislation on pollution accomplishes a constant decrease in the amount of carbon monoxide, x, over a period of ten years. During this same period liquor sales, y, have increased by a constant percentage. If the correlation coefficient between x and y is calculated, this value would point to a strongly negative relation between the level of carbon dioxide pollution and the quantity of liquor sold. However, it would be laughable to maintain that air pollution can be curbed if the public spends more money on liquor.

6 Curves that can be transformed to the linear form

In this section we study three functions of which the equation simplifies to the equation of a straight line with the aid of a suitable transformation.

6.1 The exponential curve (single-log transformation)

In the case of a linear relation between x and y, where a straight line is fitted to the data, the rate of change of $\hat{y}$ is constant. In economic and demographic data where the x values are often equally spaced, the corresponding $\hat{y}$ values form an arithmetic sequence. However, as mentioned in Chapter 4, Section 2.5, this type of data often approximates a geometric sequence rather than an arithmetic sequence. When this is the case, the relation between x and y is better described by an **exponential curve**

$$\hat{y} = ab^x$$

than by a straight line. In the literature this function is often called a growth curve, since it can be used to describe, for example, the phenomenon of economic growth over time. This relation is characterised by an almost constant relation, b, between the y values that correspond with the x values 0, 1, 2, ... By applying the rules for logarithms, it follows that the exponential curve can be expressed as follows:

$$\ell n\hat{y} = \ell n a + x\ell n b$$

If $\ell n y$, $\ell n a$ and $\ell n b$ are respectively indicated by y', a' and b', it follows that

$$\hat{y}' = a' + b'x$$

is the equation of a straight line with intercept a' and slope b'. In accordance with the analogy between the expressions in Section 2 of this chapter, the least squares values are

$$a' = \bar{y}' - b'\bar{x}$$

$$b' = \frac{\sum xy' - \frac{1}{n}\left(\sum x \sum y'\right)}{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}$$

The original form of the equation $\hat{y} = ab^x$ is then obtained by determining a and b as the antilogarithms of a' and b' respectively, i.e. $a = e^{a'}$ and $b = e^{b'}$.

EXAMPLE 6.8

In Table 6.6, y, the South African consumer price index (see Chapter 8), is given for the period 1980 to 1992. The scatter diagram for Figure 6.10 shows the scatter of points if $\ell n\, y$ is plotted against x. As these points lie more or less on a straight line, it is meaningful to fit an exponential curve to the data. The totals $\sum x$, $\sum y'$, $\sum xy'$ and $\sum x^2$ that are used in calculating a' and b' also appear in Table 6.6. Note that we select the numbers 0, 1, 2, ..., 12 as x values rather than the years 1980, 1981, 1982, ..., 1992. The use of this transformation, namely

x = year − 1980

simplifies the calculations considerably.

Table 6.6 Calculations for the fit of an exponential curve to the consumer price index for the period 1980–1992

Year t	Price index y	$\ell n\, y$ y'	$t-1980$ x	xy'	x^2	$\hat{y}$
1980	25.47	3.2375	0	0.0000	0	25.25
1981	29.33	3.3786	1	3.3786	1	28.98
1982	33.68	3.5169	2	7.0338	4	33.26
1983	37.83	3.6331	3	10.8993	9	38.17
1984	42.23	3.7431	4	14.9725	16	43.80
1985	49.06	3.8930	5	19.4652	25	50.27
1986	58.21	4.0641	6	24.3843	36	57.69
1987	67.58	4.2133	7	29.4932	49	66.20
1988	76.24	4.3339	8	34.6711	64	75.98
1989	87.48	4.4714	9	40.2427	81	87.19
1990	100.00	4.6052	10	46.0517	100	100.06
1991	115.37	4.7481	11	52.2296	121	114.83
1992	131.31	4.8776	12	58.5307	144	131.77
		52.7158	78	341.3528	650	

$\bar{x} = \frac{78}{13} = 6.0$; $\bar{y} = \frac{1\,532.0}{13} = 117.85$; $\bar{y}' = \frac{52.7158}{13} = 4.05506$

From the data in Table 6.6 it follows that

$$b' = \frac{341.3528 - \frac{1}{13}(78)(52.7158)}{650 - \frac{1}{13}(78)^2}$$

$$= \frac{25.0578}{182.0000}$$

$$= 0.13768$$

so that $b = e^{b'} = e^{0.13768}$
$= 1.1476$
while $a' = 4.05506 - 0.13768(6)$
$= 3.22898,$
so that $a = e^{a'} = e^{3.22898}$
$= 25.254$

The estimated exponential curve is therefore

$$\hat{y} = 25.254(1.1476)^x$$

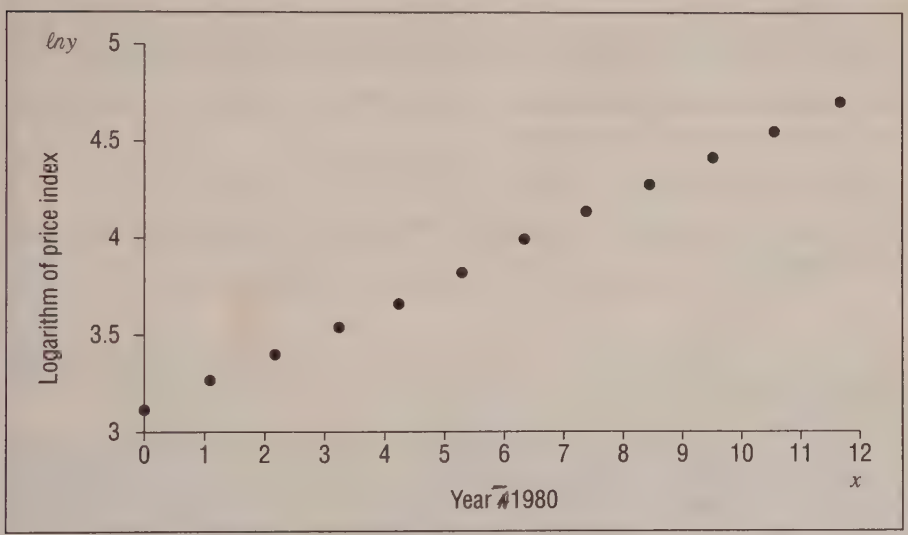


Figure 6.10 Scatter diagram of the logarithm of the consumer price index against the year for the data in Table 6.6

The $\hat{y}$ values given in the last column of Table 6.6 were calculated by substituting the different values of x in this exponential equation. With the aid of this data the coefficient of determination can now be calculated, namely

$$R^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$$

$$= 1 - \frac{7.22858}{14\,164.33571}$$

$$\approx 0.9999$$

This high value for R^2 indicates a good fit. By substituting $x = 14$ in the estimated equation, a forecast of the consumer price index for 1994 is given by

$$\hat{y} = 25.254(1.1476)^{14}$$
$$= 173.54$$

6.2 The power curve (double-log transformation)

If the observed (x, y) points are plotted and no linear relation is identified, the $(x, \ell n\, y)$ points can be represented graphically and examined for linear relation. If there is no linear relation between x and $\ell n\, y$, there may be an approximate linear relation between $\ell n\, x$ and $\ell n\, y$. When this is the case, the **power curve**

$$\hat{y} = ax^b$$

is a suitable curve for describing the relation between x and y. In logarithmic form this equation can be written as

$$\ell n\, \hat{y} = \ell n\, a + b\ell n\, x$$

If y', a' and x' respectively represent $\ell n\, y$, $\ell n\, a$ and $\ell n\, x$, we obtain the equation of a straight line

$$\hat{y}' = a' + bx'$$

with intercept a' and slope b. The least squares values (see Section 2 of this chapter) are given by

$$a' = \bar{y}' - b\bar{x}'$$

$$b = \frac{\sum x'y' - \frac{1}{n}\left(\sum x' \sum y'\right)}{\sum x'^2 - \frac{1}{n}\left(\sum x'\right)^2}$$

The coefficient a in the equation $\hat{y} = ax^b$ is determined from $a = e^{a'}$.

An application of the power function is to describe the relation between the unit price of a product (y) and the cumulative number of units (x) manufactured over a certain period. In the econometric literature this relation is known as a learning curve. The essential characteristics of learning curves are that workers learn to carry out the same tasks more quickly and effectively, managers and engineers rely on previous experience to improve the management and layout of a factory, while machinery is improved as new technology becomes available.

EXAMPLE 6.9

In columns 1 and 2 of Table 6.7 the unit price for manufacturing video machines (y) and the cumulative number (x) manufactured over a period of eight years are given respectively. Figure 6.11 is a scatter diagram of the ($\ell n\, x$, $\ell n\, y$) values for this data. A study of the figure indicates that these points lie more or less on a straight line. From this it follows that the power curve should provide a good description of the relation between x and y. In the third to the seventh columns of Table 6.7 the totals $\Sigma x'$, $\Sigma y'$, $\Sigma x'y'$, $\Sigma x'^2$ and $\Sigma y'^2$ are calculated.

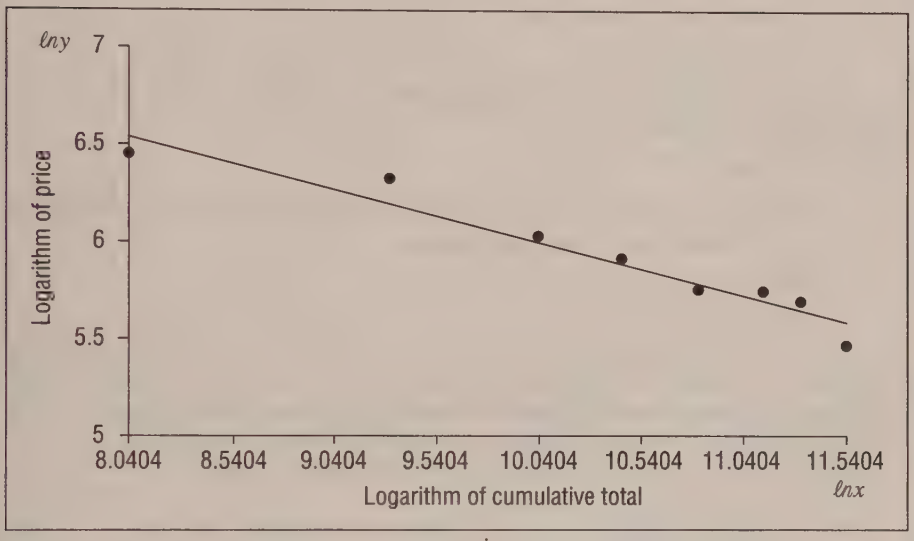


Figure 6.11 Scatter diagram of $\ell n\, y$ against $\ell n\, x$ for the data given in Table 6.7

Table 6.7 Calculations for the fit of a power curve to describe the relation between the unit price (dollars) for manufacturing video machines and the cumulative number sold

Cumulative number x	Price (dollars) y	$\ell n\, x$ x'	$\ell n\, y$ y'	$x'y'$	x'^2	y'^2	$\hat{y}$
3 104	623	8.04045	6.43455	51.73663	64.64879	41.40339	682.99
10 915	552	9.29789	6.31355	58.70270	86.45082	39.86089	487.56
22 582	407	10.02491	6.00881	60.23780	100.49879	36.10584	401.23
34 099	365	10.43702	5.89990	61.57737	108.93146	34.80879	359.27
50 216	314	10.82409	5.74939	62.23194	117.16090	33.05552	323.86
68 705	310	11.13758	5.73657	63.89152	124.04563	32.90826	297.76
83 118	295	11.32802	5.68698	64.42215	128.32396	32.34169	282.94
103 676	238	11.54903	5.47227	63.19940	133.38000	29.94575	266.67
		82.63899	47.30202	485.99950	863.44034	280.43012	

$$\bar{x}' = \frac{82.63899}{8} = 10.32987; \qquad \bar{y}' = \frac{47.30202}{8} = 5.91275$$

From the data in Table 6.7 it follows that

$$b = \frac{485.99950 - \frac{1}{8}(82.63898)(47.30202)}{863.44034 - \frac{1}{8}(82.63898)^2}$$

$$= \frac{-2.6243}{9.7902}$$

$$= -0.26805$$

If this value for b is substituted in the expression for a', we get

$$a' = \tfrac{1}{8}(47.30202) - \tfrac{1}{8}(-0.26805)(82.63899)$$

$$= 8.68268$$

The value for a is obtained as the antilogarithm of a' ($a = e^{a'} = e^{8.68168} = 5\,893.94$), so that the power curve is given as

$$\hat{y} = 5\,893.94x^{-0.26805}$$

The $\hat{y}$ values given in the last column of Table 6.7 are required for the calculation of R^2. If $x_1 = 3\,104$ it follows that, for example

$$\hat{y}_1 = 5\,893.94(3\,104)^{-0.26805}$$

$$\approx 683$$

The reader should note that the $\hat{y}$ values given in Table 6.7 could differ somewhat from those calculated with the aid of the a and b values given above. Such differences are due to rounding errors. The reader may calculate the value of $\hat{y}$ with a pocket calculator if a and b are rounded to 5 895 and -0.268 respectively.

The coefficient of determination R^2 is now calculated as follows:

$$R^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$$

$$R^2 = 1 - \frac{9\,030.77}{125\,720}$$

$$\approx 0.93$$

It thus seems that a power function is a relatively good description of the relation between the (x, y) values of Table 6.7.

6.3 The hyperbolic curve (inverse transformation)

A curve with the equation

$$\hat{y} = a + \frac{b}{x}, \qquad x \neq 0$$

is called a **hyperbolic curve**. Figure 6.12 shows the two possible forms of this curve. If $z = \frac{1}{x}$, the inverse of x, the relation

$$\hat{y} = a + bz$$

follows from the hyperbolic equation.

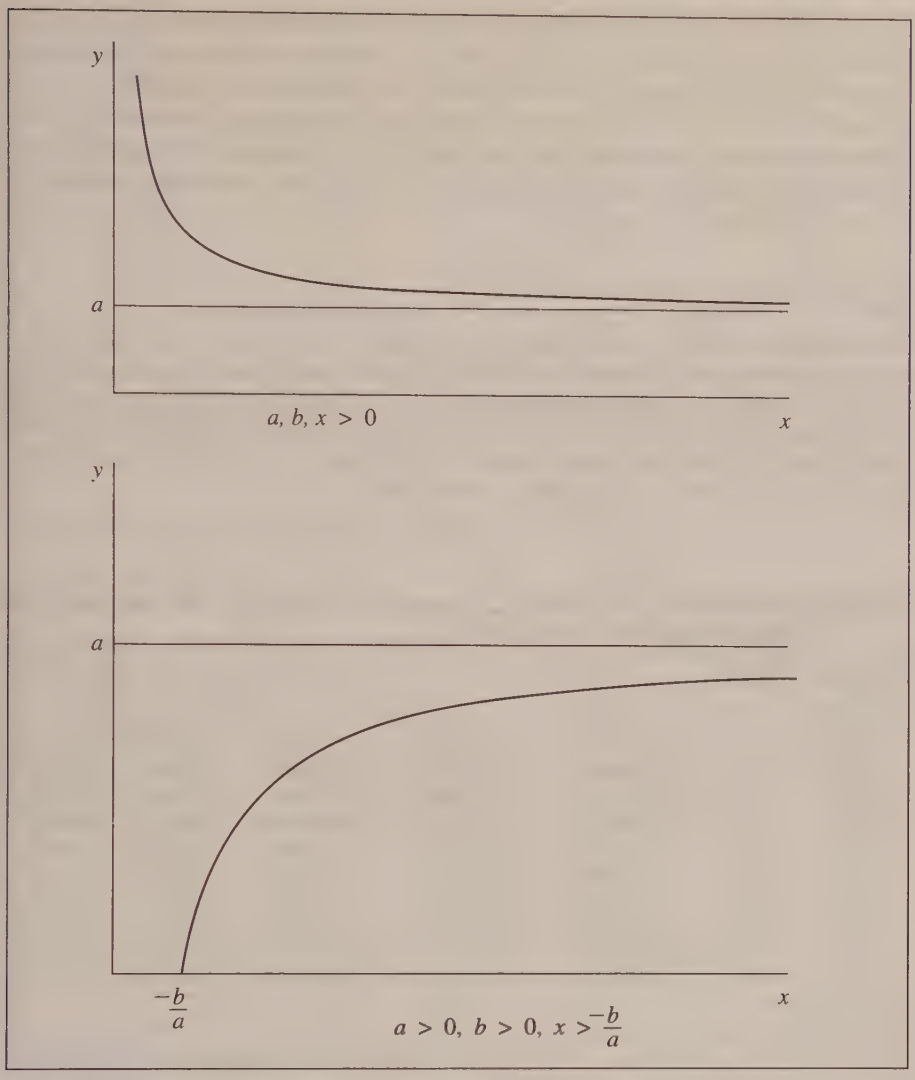


Figure 6.12 Two types of hyperbolic functions

If the relation between x and y can be described by a hyperbolic function, the relation between z and y would be linear. When none of the scatter diagrams (x, y), $(x, \ell n\, y)$ or $(\ell n\, x, \ell n\, y)$ shows a linear relation, y can be plotted against z. If the scatter diagram shows an approximately linear relation, the least squares values for a and b can be calculated from

$$a = \bar{y} - b\bar{z}$$

$$b = \frac{\sum zy - \frac{1}{n}\left(\sum z \sum y\right)}{\sum z^2 - \frac{1}{n}\left(\sum z\right)^2}$$

This hyperbolic function is often useful to describe the relation between the demand (y) for and the price (x) of a product. Another application is the relation that often exists between the stock of a specific product that a grocery shop has on hand (y) and the time elapsed since the beginning of the month (x). It follows that most consumers buy groceries at the beginning of the month and for this reason the stock does not decrease regularly, but shows a decrease as in Figure 6.12 (case $a, b, x > 0$).

EXAMPLE 6.10

The second column of Table 6.9 reflects a furniture dealer's annual demand (y) for genuine leather lounge suites over a period of eight years. The first column of this table shows the hire purchase interest rate (x) for the corresponding period. The scatter diagram of the data given in Figure 6.13 indicates that a hyperbolic curve should provide a good fit. Values for the transformation $z = \frac{1}{x}$, as well as the totals Σz, Σy, Σzy, Σz^2 and Σy^2 are given in Table 6.10.

Table 6.8 Calculations for fitting a hyperbolic curve to the demand for genuine leather lounge suites

Percentage interest x	Demand y	$z = \frac{1}{x}$	zy	z^2	y^2	$\hat{y}$
14.50	178	0.068966	12.275862	0.004756	31 684	172.5
15.50	161	0.064516	10.387097	0.004162	25 921	162.1
18.75	133	0.053333	7.093333	0.002844	17 689	136.0
22.00	115	0.045455	5.227273	0.002066	13 225	117.6
13.50	185	0.074074	13.703704	0.005487	34 225	184.4
8.75	273	0.114286	31.200000	0.013061	74 529	278.3
9.75	255	0.102564	26.153846	0.010519	65 025	250.9
16.25	157	0.061538	9.661538	0.003787	24 649	155.1
	1 457	0.584732	115.702653	0.046684	286 947	

$\bar{z} = \frac{0.5847}{8} = 0.0731; \quad \bar{y} = \frac{1457}{8} = 182.125$

From the data in Table 6.8 it follows that

$$b = \frac{115.702653 - \frac{1}{8}(0.584732)(1\,457)}{0.046684 - \frac{1}{8}(0.584732)^2}$$

$$= \frac{9.208338}{0.003945}$$

$$\approx 2\,334$$

If this value for b is substituted in the expression for a, we obtain

$$a = 182.125 - (2\,334)(0.0731)$$

$$\approx 11.51$$

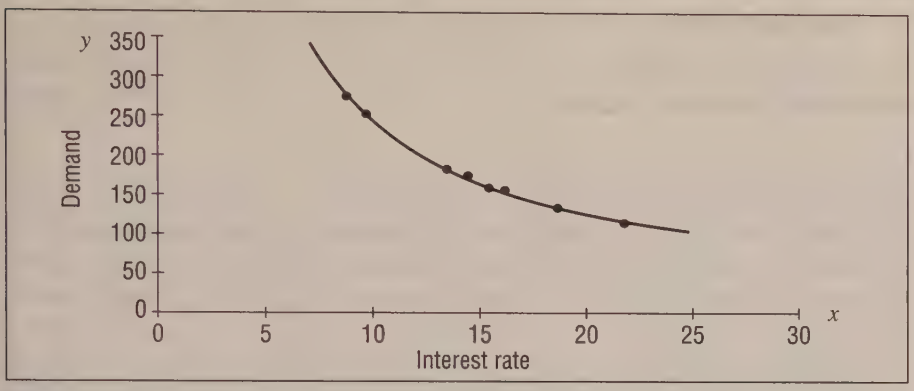


Figure 6.13 Scatter diagram of y against $z = \frac{1}{x}$ for the demand for genuine leather lounge suites given in Table 6.8

The equation of the least squares hyperbolic curve is therefore

$$\hat{y} = 11.51 + \frac{2\,334}{x}$$

The $\hat{y}$ values given in the last column of Table 6.9 are required to calculate R^2. For $x = 15.5$ it follows that

$$\hat{y} = 11.51 + \frac{2\,334}{15.5}$$

$$\approx 162.1$$

The coefficient of determination R^2 is now calculated as follows:

$$R^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$$

$$R^2 = 1 - \frac{95.676}{21\,590.875}$$

$$\approx 0.996$$

which indicates a good fit.

Note

In this section, the coefficient of determination was calculated in terms of the observed y values and the $\hat{y}$ values obtained, after having transformed back from a linear equation to the original form of the equation. If a computer package is used to fit a non-linear curve with the aid of a transformation to a linear form, the intercept and slope values of the straight line are calculated. The R^2 value that usually appears in the computer print-out provides an indication of the strength of the linear relation of the transformed data and will therefore probably differ from the R^2 value as calculated in this section.

7 The parabola

A meaningful extension of the relation $\hat{y} = a + bx$ between y and x is a **polynomial of degree q**, namely

$$\hat{y} = a + bx + cx^2 + \ldots + hx^q$$

where the degree q can now be any of the values $2, 3, \ldots, n-1$.

Although the relation between x and y is described increasingly accurately as q becomes greater, due to practical and theoretical considerations we endeavour to select a polynomial of as low a degree as possible. For this reason a **parabola**, namely

$$\hat{y} = a + bx + cx^2$$

is often used in curve fitting. For example, the relations between yield (y) and labour (x) and between production cost (y) and yield (x) are more or less parabolic.

By extending the method discussed in Section 3 of this chapter, the following three normal equations are found, from which the least squares values of a, b and c can be determined:

$$\begin{aligned} \Sigma y &= na + b\Sigma x + c\Sigma x^2 \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \\ \Sigma x^2y &= a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \end{aligned}$$

A phenomenon that is often found in the manufacturing industry is that the unit price (y) for a certain item drops as the number of items produced (x) rises. This phenomenon continues until the factory reaches full production capacity. If the demand becomes so high that more than the optimum number of items have to be produced, the unit price can rise. A possible factor contributing to this phenomenon, is that staff have to work overtime at additional remuneration to meet the increased demand. A further factor is that increased production might be possible only if money is spent on expanding the factory's facilities. The relation between unit costs (y) and number of items manufactured (x) can often be well described by a parabolic curve.

EXAMPLE 6.11

A specific spare part for motor vehicles has been manufactured locally for quite a while. The first column of Table 6.9 indicates the number of parts (in thousands) produced over this period, while the second column of the table indicates the unit price (in cents). The totals Σx, Σy, Σxy, etc. that are required for fitting a parabola to the data are calculated with the aid of the information in columns 3 to 8 of Table 6.9.

Table 6.9 Calculations for fitting a parabola

Number (× 1 000) (1) x	Price (cents) (2) y	(3) xy	(4) x^2	(5) x^2y	(6) x^3	(7) x^4	(8) y^2	(9) $\hat{y}$
2	48	96	4	192	8	16	2 304	47.86
14	36	504	196	7 056	2 744	38 416	1 296	33.24
22	29	638	484	14 036	10 648	234 256	841	25.55
26	16	416	676	10 816	17 576	456 976	256	22.33
38	11	418	1 444	15 884	54 872	2 085 136	121	15.11
46	15	690	2 116	31 740	97 336	4 477 456	225	12.35
58	13	754	3 364	43 732	195 112	11 316 496	169	11.30
74	9	666	5 476	49 284	405 224	29 986 576	81	15.65
86	31	2 666	7 396	229 276	636 056	54 700 816	961	23.22
94	34	3 196	8 836	300 424	830 584	78 074 896	1 156	30.33
106	39	4 134	11 236	438 204	191 016	126 247 696	1 521	44.06
566	281	14 178	41 228	1 140 644	3 441 176	307 618 736	8 931	

$\bar{x} = \frac{566}{11} = 51.45455; \quad \bar{y} = \frac{281}{11} = 25.54545$

The normal equations are as follows:

$$\begin{aligned} 281 &= 11a + 566b + 41\,288c \\ 14\,178 &= 566a + 41\,288b + 3\,441\,176c \\ 1\,140\,644 &= 41\,288a + 3\,441\,176b + 307\,618\,736c \end{aligned}$$

Solution of these gives:

$$a = 50.6515; \quad b = -1.4233; \quad c = 0.01284$$

from which it follows that

$$\hat{y} = 50.6515 - 1.4233x + 0.01284x^2$$

Note that the y units are in cents and the x units in thousands.

By substituting the observed x values in this equation, the $\hat{y}$ values given in the last column of Table 6.9 are now obtained. The reader is left to show that the coefficient of determination for this fit is given by $R^2 = 0.86869$.

7.1 Adjusted coefficient of determination (R_a^2)

In Example 6.9 a power curve was fitted to describe the relation between unit cost and the cumulative number of items manufactured. A scatter diagram of the raw data indicates that the relation between the (x, y) values is more or less parabolic. As in the previous example, a parabola can also be fitted to the data in Table 6.7. In this case the R^2 value works out to be 0.94. Since the R^2 value obtained with the aid of the power curve is 0.93, it would seem as though a parabola provides a better description of the relation between y and x than a power function would. However, the number of coefficients to be solved from the least squares equations (3 for the parabola and 2 for the power curve) also plays a role, so this type of comparison cannot be made directly.

Most computer software packages for fitting curves give the ordinary R^2 value as well as a so-called adjusted R^2 value. The formula according to which the latter is calculated does take the number of coefficients into account.

The **adjusted coefficient of determination**, indicated by the symbol R_a^2, is defined as follows:

$$R_a^2 = 1 - \frac{\sum (y - \hat{y})^2/(n - k)}{\sum (y - \bar{y})^2/(n - 1)}$$

where n and k respectively indicate the sample size and the number of coefficients.

For the data in Example 6.9 we can now show, for example, that

$$R_a^2 \text{ (parabola)} = 0.91, \text{ while } R_a^2 \text{ (power curve)} = 0.92$$

so, according to this data the power curve is preferred.

EXAMPLE 6.12

In the second column of Table 6.10 the gross foreign debt (y) of the RSA is given in million rand units for the period 1981 to 1989 (*South African Statistics*, 1990). A scatter diagram of the data is given in Figure 6.14.

In Table 6.10 we selected the x values in such a way that $\sum x$ and consequently also $\sum x^3$ are zero. If

$$\sum x = \sum x^3 = 0$$

is substituted in the normal equations, it simplifies to

$$\begin{aligned} \sum y &= na + c\sum x^2 \\ \sum xy &= b\sum x^2 \\ \sum x^2 y &= a\sum x^2 + c\sum x^4 \end{aligned}$$

From the second equation it follows that

$$b = \frac{\sum xy}{\sum x^2}$$

while a and c can be determined from the first and third equations. This transformation of the time scale therefore facilitates the calculation of a, b and c. Note that the x units are still years, but that the origin now corresponds with 1985.

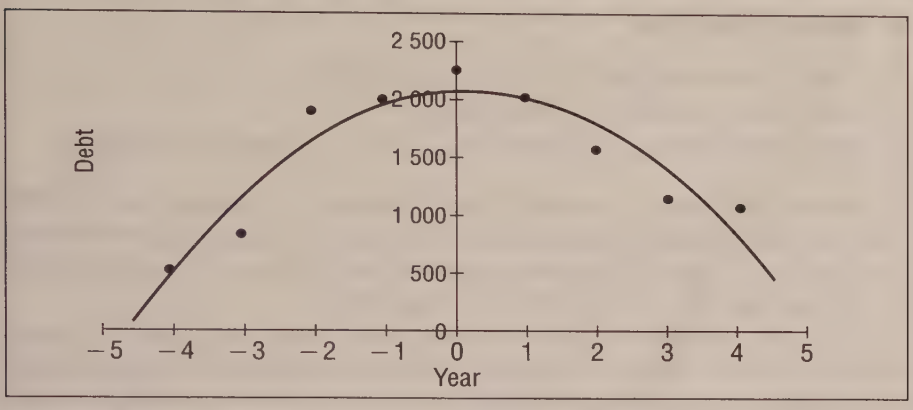


Figure 6.14 Fitting a parabola to the foreign debt data in Table 6.10

Table 6.10 Gross foreign debt (million rand) – RSA

Year	Debt	Year–1985						
	y	x	xy	x^2	x^2y	x^3	x^4	$\hat{y}$
1981	536	−4	−2 144	16	8 576	−64	256	486.6
1982	839	−3	−2 517	9	7 551	−27	81	1 155.9
1983	1 907	−2	−3 814	4	7 628	−8	16	1 644.8
1984	2 012	−1	−2 012	1	2 012	−1	1	1 953.0
1985	2 244	0	0	0	0	0	0	2 080.7
1986	2 024	1	2 024	1	2 024	1	1	2 027.9
1987	1 575	2	3 150	4	6 300	8	16	1 794.5
1988	1 133	3	3 399	9	10 197	27	81	1 380.5
1989	1 040	4	4 160	16	16 640	64	256	786.0
	13 310	0	2 246	60	60 928	0	708	

The data from Table 6.10 gives

$$b = \frac{2\,246}{60}$$
$$= 37.433$$

The two remaining equations are

$$13\,310 = 9a + 60c$$
$$60\,928 = 60a + 708c$$

Solving these equations yields $a = 2\,080.736$ and $c = -90.277$. The least squares parabola for this data is therefore (with 1985 as origin)

$$\hat{y} = 2\,080.7360 + 37.433x - 90.277x^2$$

With the aid of the data in Table 6.10 the coefficient of determination can be calculated, namely

$$R^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$$
$$= 1 - \frac{375\,766.0}{2\,970\,024.9}$$
$$= 0.873$$

8 Growth curves

The class of curves studied thus far can be classified as

(i) linear or transformable to a linear form.
(ii) polynomials of degree q. Only the parabola ($q = 2$) was discussed fully.

A characteristic that is common to both the above classes of curves is the relatively simple normal equations that are obtained by means of the method of least squares. Values of the unknown coefficients (parameters) are determined by the simultaneous solution of the sets of equations.

In this section we shall discuss three types of curves that cannot be transformed to a linear form. If these curves are fitted to data sets by means of the least squares method, normal equations that can only be solved iteratively are obtained. These curves are particularly applicable where the relation between growth (y) and time (x) has to be described – hence the name growth curves. Examples of applications are population growth, industrial growth, biological growth of man, animals, plants and other organisms, and the growth of a colony of bacteria. These curves can also be used to describe certain relations in the economy, and also in the insurance industry.

Although the calculation aspects of the fit of these curves is complex, the calculations can be done using standard computer program packages (see Chapter 1, Section 5). To understand and interpret this program output, both the functional characteristics of these curves and the meaning of the parameters should be understood.

8.1 Modified exponential curve

The equation of the **modified exponential curve** is

$$\hat{y} = a(1 - bc^x),\ a > 0,\ b > 0,\ 0 \leq c \leq 1$$

If this curve describes the relation between x and y as well as a polynomial, the modified exponential curve is usually preferred, since each of the parameters a, b and c has a specific physical meaning. The **asymptote** a represents the eventual size, the so-called 'saturation value' of the variable y. Since $b = (a - \frac{\hat{y}_0}{a})$, where $\hat{y}_0$ is the value of y when $x = 0$, the constant b is the potential growth relative to a from $x = 0$ until the saturation value is reached. The constant c is related to the growth rate. An increase in the value of c leads to a decrease in the growth rate and vice versa.

An example of a modified exponential curve is given in Figure 6.15.

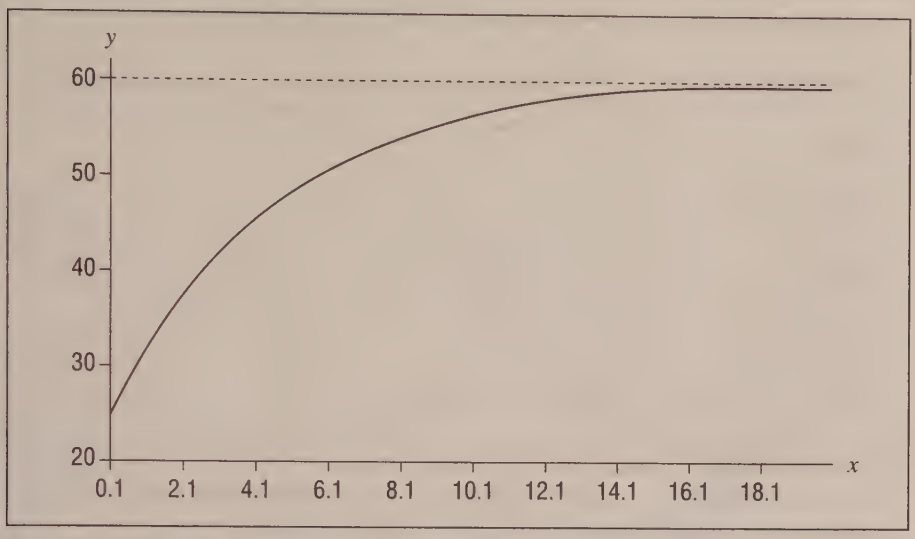


Figure 6.15 Graphic representation of the modified exponential curve $\hat{y} = 60\{1 - 0.6(0.8)^x\}$

8.2 Logistic curve

A **logistic curve** is defined by the equation

$$\hat{y} = \frac{a}{1 + bc^x}, \quad a > 0, \quad 0 \leq c \leq 1$$

in which the constants a and c have the same meaning as in the modified exponential curve and in which b is now a function of the potential growth.

Note that

$$\frac{1}{\hat{y}} = \frac{1}{a}\left(1 + bc^x\right)$$

so that a logistic relation between y and x implies a modified exponential relation between $\frac{1}{y}$ and x.

In Figure 6.16 a typical logistic curve is represented graphically. From this figure it is clear that the curve is symmetrical around its so-called inflexion point A in the form of an extended S.

8.3 Gompertz curve

The **Gompertz curve** is defined by the following equation:

$$\hat{y} = ab^{c^x}$$

By applying the rules of logarithms, it follows that

$$\ell n\, y = \ell n\, a + c^x\, \ell n\, b$$

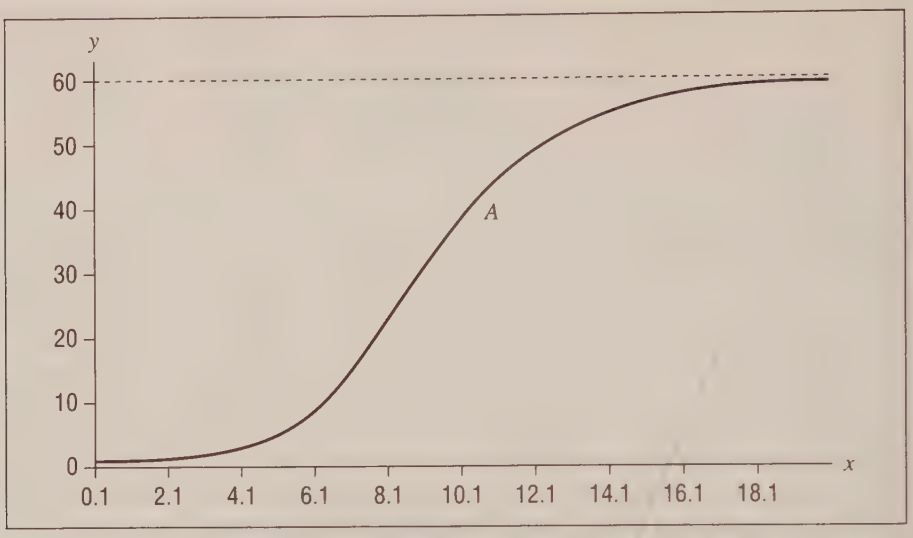


Figure 6.16 Graphic representation of the logistic curve $\hat{y} = \dfrac{60}{1 + 100(0.6)^x}$

which in the usual notation is expressed as

$$\hat{y}' = a' + b'c^x$$

This equation can be written in the form of the modified exponential curve as follows:

$$\hat{y}' = a'\left(1 - b^* c^x\right), \quad b^* = \frac{-b'}{a'}$$

Since both the logistic and the Gompertz curves can be transformed to the form of the modified exponential curve, these three curves can be regarded as belonging to the same family or class of functions.

This family is known as the **Richards family of growth curves**. Figure 6.17 represents a typical Gompertz curve. Note that, unlike the logistic curve, the Gompertz curve is not symmetrical around the inflexion point.

EXAMPLE 6.13

The data in the first two columns of Table 6.11 was collected during an investigation into the behaviour of fieldmice. A brief background description of the relevant data is as follows: seven female and six male mice were placed in a cage of 36 square metres with eight incubators. The mice were given sufficient food and water and allowed to breed freely. The births of mice were recorded daily and the mass of each mouse was determined every week. The first column of Table 6.11 shows the number of weeks since birth (x), while the second column shows the corresponding average mass in grams (y) of 40 mice.

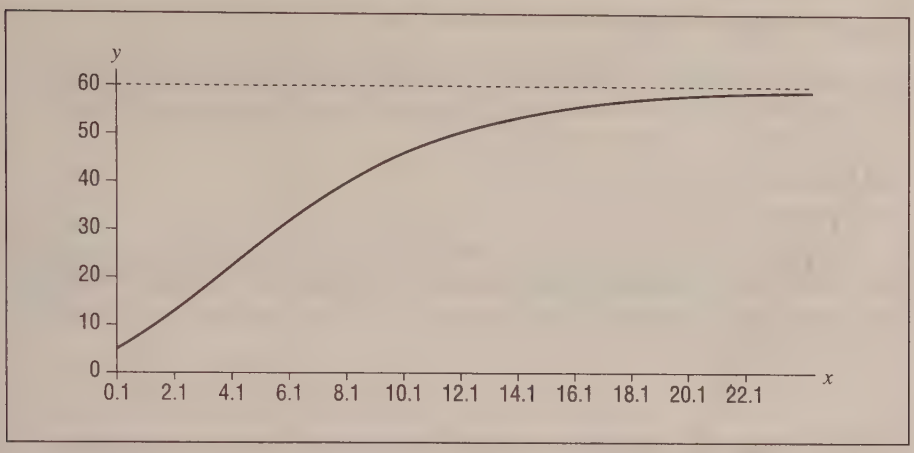


Figure 6.17 Graphic representation of the Gompertz curve $\hat{y} = 60(0.082)^{0.8x}$

By means of a suitable computer program, a logistic curve was fitted to the data. The equation of this curve is

$$\hat{y} = \frac{46.04}{1 + 13.720(0.596)^x}$$

Table 6.11 Average mass (grams) of 40 field mice for nine consecutive weeks

Week x	Average mass y	Estimated mass $\hat{y}$
3	11.33	11.77
4	17.00	16.83
5	23.00	22.62
6	28.52	28.46
7	33.33	33.65
8	37.90	37.75
9	40.48	40.71
10	42.40	42.71
11	44.43	43.99

Since the coefficient of determination $R^2 = 0.9993$, this fit is very good. Suppose we need to forecast the average mass as at 15 weeks after birth. For $x = 15$, it follows that

$$\hat{y} = \frac{46.04}{1 + 13.720(0.596)^{15}}$$

$$= 45.77 \text{ gram}$$

which is close to the saturation value of 46.04.

EXAMPLE 6.14

The first two columns of Table 6.12 give the sound intensity level data of Example 6.6. A modified exponential curve is fitted to this data. The equation obtained is

$$\hat{y} = 62.378\{1 + 0.503(0.835)^x\}$$

with the coefficient of determination

$$R^2 = 0.99$$

In Example 6.6 we showed that the coefficient of determination for a straight line fit is 0.85. This fit compares poorly to the one obtained by means of a modified exponential curve. The reader is left to show that the least squares line is given by

$$\hat{y} = 86.21 - 1.53x$$

Table 6.12 Fitting a modified curve to the sound intensity level data in Example 6.6

Distance x	Sound level y	Estimated sound level $\hat{y}$
0	95.0	93.7
2	82.0	84.2
4	77.0	77.6
6	73.5	73.0
8	71.0	69.8
10	68.5	67.5
12	66.3	66.0
14	65.0	64.9
16	63.7	64.1
18	62.5	63.6

Suppose we need to estimate the sound level at a distance of 60 metres from the sound source. By using the linear fit we find

$$\hat{y} = 86.21 - 1.53(60)$$
$$= -5.59 \text{ decibels}$$

which is a value that cannot be realized in practice. On the other hand, extrapolation by means of the modified exponential curve gives the following:

$$\hat{y} = 62.378\{1 + 0.503(0.835)^{60}\}$$
$$= 62.379 \text{ decibels}$$

Since $a = 62.379$ decibels is the so-called saturation value, it is obvious that this value can also be interpreted as the background noise (i.e. the reading that the sound level meter would register if the sound source were removed).

8.4 Determining parameters by means of the three-point method

An approximation to the least squares values of the asymptotic growth curve parameters a, b and c can be obtained by using the **three-point method**. The application of this method requires equal spacing of the x values. The data set is

divided into three groups of equal size, in which one or two of the values are omitted if the number of observations cannot be divided exactly by three.

The three-point method is subsequently illustrated using the data in Table 6.12 (see Example 6.14), which is repeated here:

		X_0			X_1			X_2			
Distance	x	0	2	4	6	8	10	12	14	16	18
Sound level	y	95	82	77	73.5	71	68.5	66.3	65	63.7	62.5
			Y_0			Y_1			Y_2		

$Y_0 = (95 \times 82 \times 77)^{\frac{1}{3}} = 84.34$

$Y_1 = (73 \times 71 \times 68.5)^{\frac{1}{3}} = 70.97$

$Y_2 = (66.3 \times 65 \times 63.7)^{\frac{1}{3}} = 64.99$

By omitting the tenth observation the data set is divided into three groups of equal size. For each group of y values the geometric mean is calculated (see Chapter 4, Section 2). Indicate the geometric means obtained by Y_0, Y_1 and Y_2 respectively. The corresponding x values, X_0, X_1 and X_2 are the middle values for the three groups of x values.

The three points (X_0, Y_0), (X_1, Y_1) and (X_2, Y_2) are substituted in the equation of the curve concerned and a, b and c are solved for the three equations. In the case of the modified exponential curve the following three formulas are derived from this method:

$$a = \frac{Y_0 Y_2 - Y_1^2}{Y_0 + Y_2 - 2Y_1}$$

$$c = \left(\frac{Y_2 - Y_1}{Y_1 - Y_0}\right)^{\frac{1}{d}}, \quad d = X_2 - X_1 = X_1 - X_0$$

$$b = \frac{a - Y_0}{ac^{x_0}}$$

From the above formulas it follows that

$a = 60.15$
$b = -0.53$
$c = 0.88$

so that

$$\hat{y} = 60.15(1 + 0.53(0.88)^x)$$

Due to the more complex nature of the formulas for a, b and c in the case of the logistic and Gompertz curves, the latter curves are preferably transformed to the modified exponential curve when the parameters concerned have to be estimated.

9 The trivariate linear model

When there is a linear relation between the dependent variable y and the explanatory variable x, this relation can be described by means of the equation $\hat{y} = a + bx$. The value of the coefficient of determination r^2 indicates whether this model is meaningful or not. If the value of r^2 is relatively low, it means that only a small percentage of the variation can be explained by x. However, this percentage can often be increased by adding additional explanatory variables.

In many practical situations it is easy to see that the result of a specific event depends on various factors. Suppose, for instance, that the valuation of a specific residence on the West Rand is of interest. The amount for which the residence is eventually valued depends on factors such as the floor areas of the main building and the outbuildings, the number of bathrooms, the size of the stand and the age of the house. The demand for a specific car model is determined by factors such as the price of the car, the price of competitive models, the guarantee period, the appearance of the car, its fuel consumption and the after-sales service. This argument indicates the necessity of using models that are formulated in terms of more than one explanatory variable. For purposes of simplicity, however, we shall discuss only trivariate data sets in this section.

Suppose there is a linear relation between the dependent variable y and two explanatory variables x and z. This relation can be described by the following equation or model:

$$\hat{y} = a + bx + cz$$

The application of the least squares method provides the following three normal equations, from which a, b and c can be solved:

$$\begin{aligned} \Sigma y &= na + b\Sigma x + c\Sigma z \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 + c\Sigma xz \\ \Sigma zy &= a\Sigma z + b\Sigma xz + c\Sigma z^2 \end{aligned}$$

From the first normal equation it follows that

$$a = \bar{y} - b\bar{x} - c\bar{z}$$

If a is eliminated from the normal equations, b and c can be solved from the two resulting equations.

EXAMPLE 6.15

In the valuation of residences it is logical to assume that the value of the land and the total floor area of the buildings are factors that play an important role. In Table 6.13 the first three columns indicate the land valuations in thousand rands, x, the total floor areas in square metres, z, and the building valuations in thousand rands, y, of ten residences.

Table 6.13 Land valuation x (R1 000), total floor area z (m^2) and building valuation y (R1 000) for ten residences

Land valuation x	Floor-area z	Building valuation y	xy	x^2	xz	zy	z^2
100.0	256	280	28 000	10 000.00	25 600	71 880	65 536
230.0	392	300	69 000	52 900.00	90 160	11 600	153 664
78.5	144	130	10 205	6 182.25	11 304	18 720	20 736
90.0	188	230	20 700	8 100.00	16 920	43 240	35 344
50.0	138	185	9 250	2 500.00	6 900	25 530	19 044
100.0	286	275	27 500	10 000.00	28 600	78 650	81 796
85.0	105	140	11 900	7 225.00	8 925	14 700	11 025
75.0	230	300	22 500	5 625.00	17 250	69 000	52 900
90.0	180	250	22 500	8 100.00	16 200	45 000	32 400
110.0	246	265	29 150	12 100.00	27 060	65 190	60 516
1 008.5	2 165	2 355	250 705	122 712.25	248 919	549 310	532 961

$\bar{x} = \frac{1008.5}{10}$; $\bar{z} = \frac{2165}{10}$; $\bar{y} = \frac{2355}{10}$

From Table 6.13 the normal equations follow, namely

$$\begin{aligned} 2\,355 &= 10a + 1\,008.5b + 2\,165c \\ 250\,705 &= 1\,008.5a + 122\,712.25b + 248\,919c \\ 549\,310 &= 2\,165a + 248\,919b + 532\,961c \end{aligned}$$

Solving these equations yields:

$$a = 100.626;\ b = -0.86477;\ c = 1.02581$$

Consequently,

$$\hat{y} = 100.626 - 0.86477x + 1.02581z$$

Table 6.14 contains the information required to calculate the coefficient of determination. The $\hat{y}$ values in column 2 were obtained by substituting the (x, z) values given in Table 6.13 in the estimated equation. Thus it follows, for example, that

$$\begin{aligned} \hat{y} &= 100.626 - 0.86477(100.0) + 1.02581(256) \\ &= 276.8 \text{ thousand rand} \end{aligned}$$

Table 6.14 Computation table for calculating R^2 with respect to the data in Table 6.13

y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$	y^2
280	276.7564	3.2436	10.5212	78 400
300	303.8464	−3.8464	14.7949	90 000
130	180.4582	−50.4582	2 246.0294	16 900
230	215.6490	14.3510	205.9518	52 900
185	198.9493	−13.9493	194.5824	34 225
275	307.5307	−32.5307	1 058.2438	75 625
140	134.8306	5.1694	26.7227	19 600
300	271.7046	28.2955	800.6325	90 000
250	207.4425	42.5575	1 811.1408	62 500
265	257.8506	7.1494	51.1145	70 225
2 355	2 355.0181	0.0	6 719.7341	590 375

$\Sigma (y - \bar{y})^2 = \Sigma y^2 - \frac{1}{10}(\Sigma y)^2 = 35\,772.5$

The value of the coefficient of determination is

$$R^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$$
$$= 1 - \frac{6\,719.73}{35\,772.5}$$
$$\approx 0.81$$

When we discussed the relation $\hat{y} = a + bx$, we mentioned that $\Sigma y = \Sigma \hat{y}$. This result also applies to the trivariate linear relation $\hat{y} = a + bx + cz$ (compare, for example, the relevant values in Tables 6.13 and 6.14). Consequently we can also show in this case that R^2 is the relation of the total variation in $\hat{y}$ that is explained by $\hat{y}$. For the current example it therefore follows that 81% of the variation in the y values can be explained by using a linear function in the explanatory variables x and z. This means that information on land valuation and floor area of buildings is expected to provide a reasonably reliable indication of the value of the residence. Suppose the stand of a specific residence in Johannesburg is valued at R100 000 and that the floor area is 262 square metres. On the basis of the sample of size 10, given in Table 6.12, it follows that the valuation of the residence would be approximately the following:

$$\hat{y} = 100.626 - 0.86477(100) + 1.02581(262)$$
$$\approx 282.91 \text{ thousand rand}$$

EXAMPLE 6.16

In Example 6.8 an exponential curve was fitted to describe the relation between the consumer price index and time. A coefficient of determination value, R^2, of 0.999 was obtained for this fit. However, it is logical to assume that there are a number of factors that affect the consumer price index. Examples of such factors are the current gold and fuel prices, the manufacturing capacity of factories, the exchange rate and expenditure by the private and public sectors on wages and salaries.

The first three columns in Table 6.15 reflect the exchange rate, x, wages and salaries of the private sector, z, and the consumer price index, y. These figures apply to the period 1980 to 1992. Although a trivariate linear model $\hat{y} = a + bx + cz$ will not necessarily give a higher R^2 value than the exponential model, this model can be used to study the effect of the explanatory variables on the consumer price.

The normal equations follow from Table 6.12, namely

$$\begin{aligned} 852.40 &= 13a + 770.5b + 1\,740.7c \\ 39\,936.22 &= 770.5a + 56\,502.19b + 81\,230.68c \\ 143\,434.52 &= 1\,740.7a + 81\,230.68b + 294\,568.23c \end{aligned}$$

Solving these equations provides

$$a = 6.4387;\ b = -0.0436;\ c = 0.4609$$

Consequently,

$$\hat{y} = 6.4387 - 0.0436x + 0.4609z$$

Table 6.15 Exchange rate, x, private sector wages and salaries, z, and consumer price index, y, for the period 1980 to 1992

x	z	y	xy	x^2	xz	zy	z^2
109.3	48.3	25.5	2 787.15	11 946.49	5 279.19	1 231.65	2 332.89
104.8	58.3	29.3	3 070.64	10 983.04	6 109.84	1 708.19	3 398.89
90.4	69.0	33.7	3 046.48	8 172.16	6 237.60	2 325.30	4 761.00
91.6	78.5	37.8	3 462.48	8 390.56	7 190.60	2 967.30	6 162.25
72.9	89.4	42.2	3 076.38	5 314.41	6 517.26	3 772.68	7 992.36
49.4	100.0	49.1	2 425.54	2 440.36	4 940.00	4 910.00	10 000.00
42.3	114.5	58.2	2 461.86	1 789.29	4 843.35	6 663.90	13 110.25
43.7	131.8	67.6	2 954.12	1 909.69	5 759.66	8 909.68	17 371.24
37.8	153.8	76.2	2 880.36	1 428.84	5 813.64	11 719.56	23 654.44
34.3	178.9	87.5	3 001.25	1 176.49	6 136.27	15 653.75	32 005.21
33.1	208.8	100.0	3 310.00	1 095.61	6 911.28	20 880.00	43 597.44
31.1	239.9	115.4	3 588.94	967.21	7 460.89	27 684.46	57 552.01
29.8	269.5	129.9	3 871.02	888.04	8 031.10	35 008.05	72 630.25
770.5	1 740.7	852.4	39 936.22	56 502.19	81 230.68	143 434.52	294 568.23

$\bar{x} = \frac{770.5}{13} = 59.269; \quad \bar{z} = \frac{1740.7}{13} = 133.900; \quad \bar{y} = \frac{852.4}{13} = 65.569$

From the above equation it follows that an increase of one unit in the exchange rate can be expected to cause a drop of 0.0436 in the consumer price index. This assumption presupposes that wages and salaries in the private sector remain unchanged. Similarly, it follows that if the exchange rate remains unchanged, the consumer price index can be expected to rise by 0.4609 if the value of z is increased by one unit (million rand).

Table 6.16 contains the information required to calculate the coefficient of determination. The $\hat{y}$ values in column 2 were obtained by substituting the (x, z) values given in Table 6.15 in the estimated equation. Thus, for example, it follows that

$$\hat{y}_1 = 6.4387 - 0.0436(109.3) + 0.4609(48.3)$$
$$= 23.93$$

Table 6.16 Computation table for calculating R^2 with respect to the data in Table 6.15

y	$\hat{y}$	$(y - \hat{y})$	$(y - \hat{y})^2$	y^2
25.5	23.93	1.57	2.4649	650.25
29.3	28.74	0.56	0.3136	858.49
33.7	34.30	–0.60	0.3600	1 135.69
37.8	38.63	–0.83	0.6889	1 428.84
42.2	44.46	–2.26	5.1076	1 780.84
49.1	50.37	–1.27	1.6129	2 410.81
58.2	57.37	0.83	0.6889	3 387.24
67.6	65.28	2.32	5.5824	4 569.76
76.2	75.68	0.52	0.2704	5 806.44
87.5	87.40	0.10	0.0104	7 656.25
100.0	101.23	–1.23	1.5129	10 000.00
115.4	115.65	–0.25	0.0625	13 317.16
129.9	129.35	0.55	0.3025	16 874.01
852.4	852.40	0	18.7779	69 875.78

$\Sigma(y - \bar{y})^2 = \Sigma y^2 - \frac{1}{n}(\Sigma y)^2 = 13\,984.568$

The value of the coefficient of determination is

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$
$$= 1 - \frac{18.3447}{13\,984.568}$$
$$= 0.9987$$

It therefore follows that 99.87% of the variation in the y values can be explained with the aid of a linear function in the explanatory variables x and z.

Notes

(a) If there is a linear relation between the variables x, y and z, all the observation points should lie more or less on a plane. The fitted equation actually describes a plane with a as the intercept of the plane on the y axis. If z remains constant, b indicates the change in y for a unit increase in x. On the other hand, c can be regarded as the change in y for a unit increase in z at a constant value of x.

(b) Suppose that the relation between two variables x and y can be described by means of a parabola. The equation $\hat{y} = a + bx + cx^2$ can alternatively be written as $\hat{y} = a + bx + cz$, with $z = x^2$. The theory discussed in this section therefore applies directly to the fit of a parabola. In such a case, R^2 can be regarded as the relation of the variation in y that is explained by a linear function in x and x^2.

(c) Where a linear relation between more than three variables is being studied, the normal equations can be written as analogues of those with two and three variables. However, since it is time-consuming to solve these normal equations, it is better to avail oneself of existing computer program packages (see Chapter 1, Section 5).

(d) Where there is a linear relation between three variables x, y and z, it is often important to establish the correlation between, for example, x and y, once the linear influence of the variable z has been removed from both. The strength of this type of relation is measured by the partial correlation coefficient. Thus, for instance, the partial correlation coefficient between x and y, after removal of the linear influence of z, is defined by

$$r_{xy.z} = \frac{r_{xy} - r_{xz} r_{yz}}{1 - r_{xz}^2 \sqrt{1 - r_{yz}^2}}$$

In the above expression r_{xy}, r_{xz} and r_{yz} indicate the ordinary correlation coefficients between x and y, x and z and between y and z. Similarly, the other two partial correlation coefficients are given by

$$r_{yz.x} = \frac{r_{yz} - r_{xy}r_{xz}}{\sqrt{1 - r_{xy}^2}\sqrt{1 - r_{xz}^2}}$$

$$r_{xz.y} = \frac{r_{xz} - r_{xy}r_{zy}}{\sqrt{1 - r_{xy}^2}\sqrt{1 - r_{zy}^2}}$$

(e) If there is a linear relation between more than two variables, the positive square root of R^2, namely R, is called the multiple correlation coefficient. If y represents the dependent variable, it can be proven that R is no more than the ordinary correlation coefficient between y and $\hat{y}$.

Exercises

1 Consider the first ten pairs of observations of values (x) and selling prices (y) with respect to residences in Durban (Appendix A, Data Set 1, columns 4 and 3).
   (a) Draw a scatter diagram of the data; and
   (b) fit a least squares straight line to the data.

2 Repeat Exercise 1 for residences 1 to 10 in Johannesburg.

3 (a) Represent the fitted curves with respect to Exercises 1 and 2 graphically on the same axis. What conclusions can you derive from this representation?
   (b) Calculate the correlation coefficient r for both data sets.

4 Consider Chapter 5, Exercise 38. The mining company is interested in the correlation between the lead content and the zinc content in twelve samples. Draw a scatter diagram of the lead content against the zinc content and calculate the correlation coefficient r.

5 Consider the first ten pairs of observations of the selling prices (y) and the floor area (x) of residences in Durban (Appendix A, Data Set 1, columns 3 and 5).
   Determine whether a straight line would fit the data by:
   (a) drawing a scatter diagram; and
   (b) calculating the correlation coefficient r.

6 (a) Fit an exponential curve to the data in Table 6.2.
   (b) Calculate the coefficient of determination R^2.

7 (a) Fit a power curve to the data of Table 6.6. Use the numbers 0, 1, 2, ..., 12 for x values to coincide with the years 1980, 1981, ..., 1992.
   (b) Calculate the coefficient of determination R^2 and compare it with the value obtained in Example 6.8.

8 Repeat Exercise 7 by fitting a hyperbolic curve to the data.

9 (a) Fit a parabola to the data of Example 6.6 and give a graphic representation of the curve obtained for the values $x = 0$ to $x = 36$.
(b) Can the curve be used for extrapolation purposes?
(c) Calculate the coefficient of determination, R^2.

10 Consider the following data, namely, the number of divorces per thousand members of a particular group in the RSA for 1971 – 1990.
(a) Draw a scatter diagram of the data for the period 1971 – 1980.
(b) Fit a parabola to the data for the period 1971 – 1980 (take 1975 as origin and the x units as years).
(c) Could the parabola be used to extrapolate for 1990?
(d) Compare your answer to (c) with the divorce rate as recorded in 1990.

Year x	Divorce rate y	Year x	Divorce rate y
1971	2.10	1981	3.71
1972	2.11	1982	3.78
1973	2.18	1983	3.51
1974	2.39	1984	3.65
1975	2.53	1985	3.48
1976	2.52	1986	3.75
1977	2.26	1987	3.72
1978	2.60	1988	3.71
1979	3.11	1989	3.72
1980	3.65	1990	3.97

11 The table given below indicates the total number of licensed vehicles (millions) in South Africa for the period 1966 – 1991 (source: *S.A. Statistics*).

Year x	Number of vehicles y	Year x	Number of vehicles y
1966	1.854	1982	4.769
1968	1.981	1983	4.944
1970	2.344	1984	5.225
1972	2.655	1985	5.348
1974	3.152	1986	5.454
1976	3.507	1987	5.528
1978	3.572	1988	5.581
1979	3.987	1989	5.780
1980	4.126	1990	5.952
1981	4.430	1991	6.113

(a) Draw a scatter diagram of the data.
(b) Fit a least squares line to the data and calculate R^2.
(c) Fit an exponential curve to the data and calculate R^2.
(d) Estimate the expected number of vehicles for 1995 and for 2000 using both the fitted equations. Comment.

12 A works engineer is of the opinion that y, the number of defective items produced hourly, is directly proportional to x, the speed (revolutions per minute) of the lathe on which the items are produced. The following is a table of ten random observations each representing one hour.

y	9	13	6	11	10	8	7	9	12	5
x	233	148	181	265	231	208	202	258	273	165

(a) Fit a least squares line to the data.
(b) Calculate the correlation coefficient.
(c) Is the works engineer's opinion valid?

13 In the table below we are given the experimental values for the pressure p of a given gas mass at different values v of the volume. According to thermodynamics $pv^a = b$, where both a and b are constants. Determine the values of a and b and estimate p for $v = 150$.

Volume v	54.3	61.8	72.4	88.7	118.7	194.0
Pressure p	61.2	49.5	37.6	28.4	19.4	11.1

14 The following data concerns the demand for a product (thousands) and its price (rand) in five different marketing areas.

Price (in rand)	10	12	14	16	18
Demand	120	89	56	41	22

(a) Draw a scatter diagram of the data.
(b) Fit a parabola to the data and estimate the demand if the price is R15. (Clue: the normal equations can be simplified by the appropriate transformation of the price values.)
(c) Would it be expedient to estimate the demand from the fitted parabola when the price is R40? Motivate your answer.

15 (a) Fit a power curve to the following data:

y	50.0	60.0	70.0	80.0	90.0	100.0
x	64.7	51.3	40.5	25.9	7.8	6.7

(b) Calculate R^2 for the fitted curve.

16 In a survey on a number of new motor vehicles with regard to engine capacity (cc) and fuel consumption (ℓ/100 km) at 120 km/h, the following information was collected:

Engine capacity x	**Fuel consumption** y
1 272	7.32
1 595	8.10
1 971	9.11
2 788	11.68
3 453	13.24

(a) Fit a hyperbola to the data.
(b) Using (a) estimate the fuel consumption at 120 km/h of a motor vehicle with an engine capacity of 2 500cc.

17 During the first six years of its existence, the net profit (R1 000) of a company is as follows:

Year	1	2	3	4	5	6
Profit	112	149	238	354	580	867

(a) Graphically test to see whether an exponential curve can be fitted to the data.
(b) Fit an exponential curve to the data.
(c) Calculate a measure of the strength of the relation between the years in existence and the net profit.

18 A firm leasing duplicating machines envisages using the age of a machine to estimate the maintenance cost per machine. It has the particulars of the maintenance costs (R1 000) of 7 machines from the date of purchase.

Age	1	1.7	2	2.5	2.8	3.1	3.4
Maintenance	0.5	0.6	1.2	2.0	2.6	2.9	3.3

(a) Identify the dependent and explanatory variable.
(b) Draw a scatter diagram.
(c) Fit a straight line to the data.
(d) In your opinion, is the fitting satisfactory for the purpose for which the firm wants to use the curves?

19 After three months of training, six factory workers are chosen at random to determine their competence (on the basis of a test) and their productivity (on the basis of the number of items produced within a set time).

Factory worker	Competence (x)	Productivity (y)
1	9	23
2	17	35
3	20	29
4	19	33
5	20	43
6	23	32

(a) Find the equation for the least squares line to be used to estimate productivity on the basis of competence.
(b) What is the expected productivity of a worker whose competence value is 20?
(c) Determine the correlation coefficient and interpret the obtained value.

20 The manufacturers of dishwashing machines need to determine whether there is any relation between monthly sales (hundreds) and the number of advertisements televised monthly for their product.

Sales y	8.4	5.2	7.1	10.0	12.9	11.5	14.4	13.5
Number of advertisements x	11	6	8	9	12	15	14	13

(a) Draw a scatter diagram of the data.
(b) Determine the least squares line of y given x and represent this line graphically on the axes used in (a).
(c) Calculate a suitable measure of the linear relation between the sales and the number of televised advertisements per month.
(d) What percentage of variation in the y values is explained by the independent variable x?

21 Consider Chapter 5, Exercise 39.
(a) Calculate the average amount in arrears per income category.
(b) Calculate the class mark for each income group.
(c) Draw a scatter diagram for the amount in arrears (y) against income.
(d) Fit a suitable curve to the data.
(e) Calculate R^2.
(f) What conclusions can you draw from your calculations?

22 The manager of a motor car sales company is of the opinion that there is a linear relation between a salesman's ability, on the one hand, to sell new motor cars, and his powers of persuasion and his professionalism on the other. To validate his statement, ten salesmen are chosen at random and subjected to testing of their verbal powers of persuasion and their professionalism. The results are given in the following table:

Salesman	Average number of cars sold per month	Verbal powers of persuasion	Professionalism
	y	x	z
1	15	80	70
2	6	50	55
3	11	65	60
4	12	70	66
5	14	85	75
6	13	75	72
7	8	65	57
8	7	55	58
9	10	68	69
10	9	62	67

(a) Fit a curve of the form $\hat{y} = a + bx + cz$ to the data.
(b) Calculate R^2.

23 The manager of a large business concern would like to be able to predict accurately the number of days per year that workers will be absent. The data below represents x, a combined personality and fitness count of each of ten workers and y, the number of days of absenteeism of these workers in one year.

x	90	70	59	74	83	90	86	77	55	73
y	1	3	6	4	3	1	2	4	8	5

(a) Fit a curve of the form $\hat{y} = a + bx$ to the data.

(b) Calculate and interpret the correlation coefficient.
(c) Estimate the number of days a worker, whose test score is 70, can be expected to be absent.

24 The following table reflects the buying power of the rand in different years compared to a rand (100c) in 1910. (*RSA statistics in brief*, Central Statistical Service)

Year	Buying power	Year	Buying power
1910	100	1985	5
1930	72	1988	3
1950	47	1989	3
1970	25	1990	2
1975	16	1991	2
1980	9		

Fit a modified exponential curve to the data and predict the buying power of the rand in 1995 and 2000 relative to the buying power in 1910.

25 Fit a logistic curve to the data of Exercise 24, but assume that $a = 1$ (clue: given $a = 1$ it follows that $\hat{u} = (\frac{1}{y} - 1) = bc^x$, an exponential curve).

26 The following table reflects the private consumer expenditure in RSA (thousand million rand) (*RSA statistics in brief*, Central Statistical Service)

Year	Expenditure
1960	3.32
1980	30.92
1982	44.13
1983	50.26
1984	57.62

Fit an exponential curve to the data and predict the expenditure in 1985.

27 Fit a straight line to the data in Example 6.4 so that it passes through the origin. Interpret the least squares value of the parameter b.

28 Use the data in Appendix A, Data Set 5. Let y = amount of electricity (kWh) used weekly and x = the amount of beer produced weekly.
(a) Draw a scatter diagram of the data for both fiscal years. Use different symbols for the different years.
(b) Fit a suitable curve to the data for each year.
(c) Calculate R^2 for both data sets.
(d) How does the data for the two years compare?

29 Oxen are fed in a feeding-pen experiment and the correlation between the cumulative amount of fodder (kg) consumed is measured weekly. The data for two of the oxen is as follows:

Feeding days x	Cumulative amount of fodder (kg) Ox 1	Ox 2
0	0	0
8	15	40
15	71	105
22	132	170
29	199	236
36	269	307
43	328	369
50	391	432
57	455	498
64	531	575
71	570	620
78	682	733

(a) Draw a scatter diagram of the data and use a different symbol for each ox.
(b) Fit suitable curves between the cumulative fodder consumption and the number of feeding days.
(c) Calculate R^2 for each ox.

30 In a feeding-pen experiment the body mass (kg) of four oxen and the time period of the experiment is determined as follows:

Time	Body mass Ox 1	Ox 2	Ox 3	Ox 4
0	181	198	189	189
8	194	185	196	193
15	198	198	202	191
22	205	209	213	205
29	217	218	219	219
36	227	214	219	205
43	241	238	233	224
50	248	238	248	237
57	258	266	258	244
64	263	272	259	248
71	272	282	261	254
78	272	288	263	264
85	290	298	285	280
92	299	312	275	296
99	303	322	301	307
107	304	330	308	323
114	312	332	311	332

(a) Calculate the average body mass of the four oxen for each point in time and draw a scatter diagram of the average against time.
(b) Fit a suitable curve to the averages.
(c) Calculate R^2.
(d) Estimate the mean mass for the oxen at 120 and 150 days.

31 At a coal power station in Zimbabwe the ash content of the coal must be within certain limits. The ash content of the coal must be determined before it is delivered to the power station. The old method for doing this is time-consuming and often fails to provide the information in time. A new and faster method has been developed to determine the ash content of the coal. However, there is still some doubt as to the accuracy of the method. A number of coal samples are taken and the ash content is determined by both the old method and the new faster method.

Observation	Ash content	
	Old method	New method
1	20.5	20.9
2	22.0	22.2
3	22.4	22.5
4	22.5	22.8
5	22.6	23.6
6	23.2	23.7
7	23.3	24.1
8	23.4	24.1
9	23.8	24.2
10	23.9	24.2
11	24.3	24.4
12	24.3	24.6
13	24.3	24.6
14	24.7	24.8
15	24.7	25.1
16	25.0	25.2
17	25.0	25.4
18	25.2	25.4
19	25.2	26.0
20	25.2	26.2
21	25.6	26.6
22	26.4	27.0
23	26.9	27.0
24	27.2	28.0
25	27.3	29.7
26	29.4	29.9

Data obtained from MBL group BUL00193

(a) Draw a scatter diagram to compare the old method to the new method.
(b) Fit a linear regression line between the old and the new method.
(c) Calculate the correlation coefficient.
(d) Can the new method be used to replace the old method?

32 It is common knowledge that there is a relationship between the price of oil and the price of petrol. The monthly average price of oil in American dollar per vat in 1987 prices and the average price of 98 octane petrol in American cent per litre in 1987 prices are given for the period January 1988 to December 1990.

Month	Oil price (x)	Petrol price (y)
January 1988	14.79	17.28
February	14.36	16.96
March	12.89	16.97
April	14.36	17.17
May	14.01	17.91
June	12.68	17.71
July	14.01	16.15
August	12.41	15.65
September	11.08	15.48
October	9.75	15.09
November	10.14	15.08
December	11.52	16.53
January 1989	13.08	16.15
February	13.06	14.69
March	14.12	17.04
April	15.05	18.72
May	13.56	20.65
June	13.52	19.95
July	14.23	18.63
August	13.46	18.04
September	13.98	17.92
October	14.29	17.93
November	14.35	17.91
December	14.94	18.08
January 1990	14.85	18.61
February	14.39	19.22
March	13.38	19.85
April	12.76	19.85
May	12.84	19.26
June	11.90	17.87
July	13.25	18.25
August	20.48	26.10
September	25.32	32.17
October	26.98	32.04
November	23.84	26.10
December	19.21	26.64

(a) Draw a scatter diagram of the petrol price against the oil price.
(b) Fit a linear regression line to the data and plot it on the scatter diagram.
(c) Estimate the petrol price in 1987 prices if the average oil price is \$15 per vat.
(d) Estimate the petrol price in 1987 prices if the average oil price is \$25 per vat.
(e) Calculate the correlation between petrol price and oil price.

33 Consider the data in Exercise 32.
(a) Draw a scatter diagram of the oil price against time. Use the numbers 1, 2, …, 36 as x values to correspond with January 1988 to December 1990.
(b) Fit a power curve to the oil price.
(c) Estimate the oil price for January 1991 and January 1995.
(d) Calculate the coefficient of determination R^2 for this fit.

34 Consider the data in Exercise 32.
(a) Draw a scatter diagram of the oil price against time. Use the numbers 1, 2, …, 36 as x values to correspond with January 1988 to December 1990.
(b) Fit a parabola to the oil price.
(c) Estimate the oil price for January 1991 and January 1995.
(d) Calculate the coefficient of determination R^2 for this fit.

35 Consider the data in Exercise 32.
(a) Draw a scatter diagram of the oil price against time. Use the numbers 1, 2, …, 36 as x values to correspond with January 1988 to December 1990.
(b) Fit a straight line to the oil price.
(c) Estimate the oil price for January 1991 and January 1995.
(d) Calculate the coefficient of determination R^2 for this fit.

36 Consider the data in Exercise 32.
(a) Let Y be the petrol price, X the oil price and Z time. Use the numbers 1, 2, …, 36 as z values to correspond with January 1988 to December 1990.
(b) Fit the equation $\hat{y} = a + bx + cz$ to the data.
(c) Estimate the petrol price for January 1991 and for January 1995 if the oil price is \$20 per vat.
(d) Calculate the coefficient of determination R^2 for this fit.

37 Consider the turnover and number of clients in Appendix A, Data Set 4.
(a) Draw a scatter diagram of the turnover against the number of clients for April.
(b) Fit a suitable curve to the data and plot it on the scatter diagram.
(c) Estimate the turnover if there are 100 clients in the restaurant.
(d) Calculate the relation between the turnover and the number of clients.

38 Repeat Exercise 37 for the month of May.

39 A gynaecologist frequently uses sonar to monitor the development of the unborn foetus. An important measurement, for example, is the biparietal diameter in millimeters, that is, the diameter of the head of the foetus. Sonar scans have been done for a large group of patients and P_{10}, the median and P_{90} of the scans have been determined for each age group.

Age in weeks	P_{10}	P_{50}	P_{90}
13	21.0	22.0	23.0
14	25.0	26.0	28.3
15	28.6	31.0	34.0
16	34.0	35.0	37.0
17	37.0	38.0	40.0
18	40.0	41.0	43.0
19	43.0	45.0	47.0
20	45.7	47.0	50.0
21	49.0	51.0	53.0
22	52.0	54.0	56.0
23	55.0	57.0	59.0
24	58.0	60.0	62.0
25	61.0	62.0	65.0
26	64.0	66.0	68.0
27	67.0	69.0	71.0
28	70.0	72.0	74.0
29	71.0	75.0	77.0
30	75.0	77.0	79.0
31	78.0	80.0	82.0
32	79.0	82.0	84.0
33	81.0	85.0	86.0
34	83.0	86.0	89.0
35	84.0	88.0	90.0
36	86.0	89.0	91.0
37	87.0	90.0	93.0
38	88.0	92.0	94.0
39	89.3	94.0	96.0
40	91.0	95.0	98.0

(a) Draw a scatter diagram of the data.
(b) Fit the Gompertz curve to each percentile.
(c) Calculate R^2 for each curve.

40 The management of an airline service suspects that there is a direct relationship between advertising expenditure and the number of passengers preferring to fly with the airline. The monthly expenditure on advertisements and the number of passengers for the last twelve months were used.

Advertising expenditure (R1 000)	Passengers (1 000)
11.2	16.8
13.3	19.0
8.5	12.8
16.9	24.3
11.3	16.4
15.9	21.8
11.0	15.4
15.1	21.7
20.3	25.3
10.5	18.0
12.6	16.6
14.0	19.5

(a) Use a linear regression model to predict the number of passengers if R25 000 was spent on advertising.
(b) Calculate the correlation coefficient between advertising expenditure and the number of passengers.
(c) What is the percentage of variation explained by the fitted regression model?

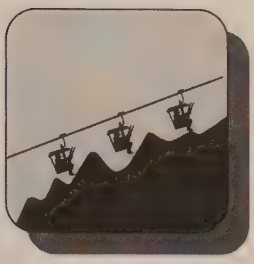
41 A marketing manager has called in a consultant to devise a multiple regression model to predict sales y (R1 000) with advertising expenditure x (R1 000) and percentage market share z.

Sales y	Advertising expenditure x	Percentage market share z
10.8	4.3	26.2
12.6	4.0	32.2
8.3	4.1	17.3
9.2	4.6	16.7
11.1	5.5	18.9
10.9	4.5	13.2
7.9	3.3	14.4
11.6	2.8	27.1
8.2	2.6	20.8
9.0	3.1	22.0
13.3	4.3	33.3
7.2	2.6	15.0

(a) Fit a curve $\hat{y} = a + bx + cz$ to the data.
(b) Calculate the estimated sales y for each observation pair (x, z).
(c) Calculate R^2.

CHAPTER 7

Time series



1 Introduction

In this chapter we shall examine bivariate data sets $\{(x_i, y_i)\}$, in which the variable x indicates time. A data set that is dependent on the time at which the data was collected is called a **time series**. The x values or **time points** $x_1, x_2, \ldots, x_n$ are chronological; x_1 is the furthest in the past and x_n is the most recent time point. The x values are usually chosen in such a way that the difference between consecutive points is equally spaced. Consequently, time series data is collected on, for example, a monthly, quarterly or annual basis. The variable y depends on the time and the observed values $y_1, y_2, \ldots, y_n$ correspond with the points $x_1, x_2, \ldots, x_n$.

A **time series** can briefly be defined as follows:

> A **time series** is a collection of observations of a phenomenon or variable which is arranged in a chronological sequence.

Since time series provide important economic and demographic information, they are published regularly by official organizations. Appendix D contains 22 typical time series which we have called TR1, TR2, ..., TR22. A brief description of each is given on the first page of the appendix. These time series contain, *inter alia*, information on internal and foreign trade, agricultural production, birth rates, road accidents and mining.

Figure 7.1 is a graphic representation of the annual rainfall for Cape Town over the period 1965 to 1991 (see Appendix D, TR3). Note that in this type of representation, in contrast with the scatter diagrams discussed in Chapter 6, the consecutive points are usually connected by straight lines.

Forecasts based on observed data are usually the main objective of time series analysis. Such forecasts or future projections are essential for meaningful planning. An observed time series will usually contain one or more of four kinds of variations or so-called **movement components**. Each of these movement components can be attributed to a particular cause. A brief description of each movement component now follows.

Section 1.5 contains an exposition on how these four movement components can be combined to form the **multiplicative model**.

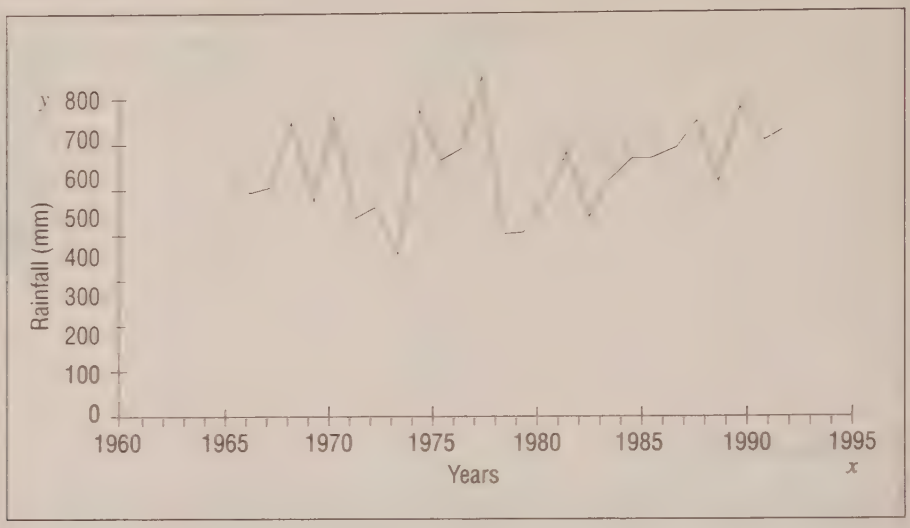


Figure 7.1 Annual rainfall (mm) for Cape Town, 1965 to 1991

1.1 The long-term movement (*L*)

Most time series show a general trend or tendency which can be observed only in the long term. Figure 7.2 is a graphic representation of the gross value of wheat production for the period 1971 to 1991 (see Appendix D, TR4). This series shows a gradual increase in gross value. However, some series may show a decrease or may, in the long term, remain constant.

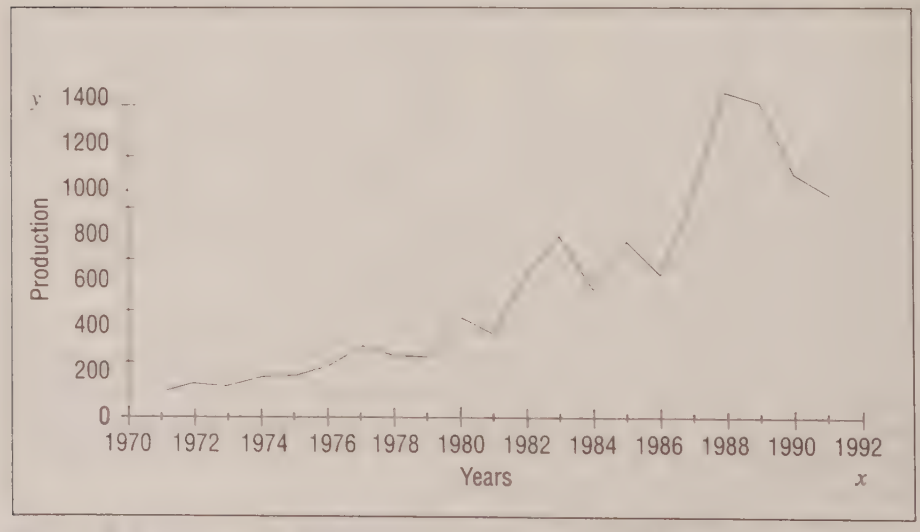


Figure 7.2 Gross value of wheat production (million rand), 1971 to 1991

1.2 Seasonal fluctuation (*S*)

Some time series show identical variation patterns that are repeated, for example, on a monthly or annual basis. Figure 7.3 is a graphic representation of the grain sorghum stocks at the end of each month for the period January 1989 to February 1992 (see Appendix D, TR2.2).

This figure clearly shows that amid the gradual decrease in stock, a pattern emerges that is repeated annually. The stock is at its highest level during July each year while it reaches a low during the months of April and May. A movement component is classified as a seasonal fluctuation only when the period over which the pattern is repeated is one year or less.

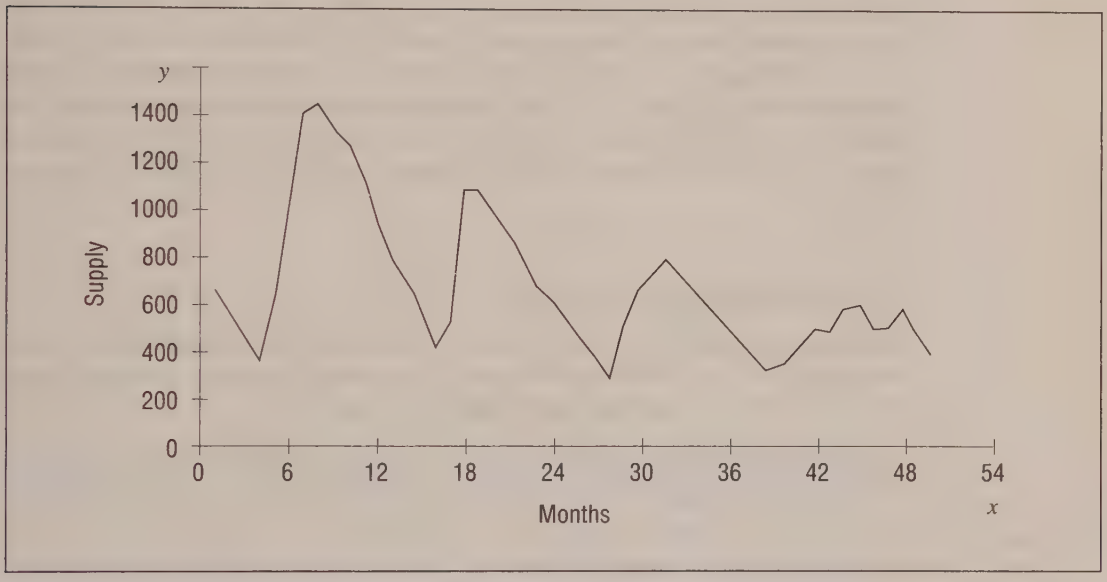


Figure 7.3 Month-end grain sorghum stocks (1 000 tons) for January 1989 to February 1992

1.3 The cyclical movement or business cycle (*C*)

In many time series there is a long-term swing above and below the general trend. The period of this component is not necessarily constant, but usually varies from a few years to several decades. In the case of economic data which has been observed over many years, the cyclical movement can be identified as a so-called **business cycle**. The four phases in a business cycle are prosperity, recession, depression and recovery. When an observed time series extends over a period of less than five years, it is difficult to distinguish between the trend and the business cycle.

1.4 Irregular variation (*I*)

If a time series is represented graphically, the sequence is often irregular and it resembles unevenly distributed saw-teeth. This **irregular** movement can be ascribed to observation errors and to incidental phenomena such as hail damage, earthquakes, strikes, elections or wars. The annual rainfall represented in Figure 7.1, is a time series that is almost entirely comprised of an irregular component.

Our main purpose in analysing time series is to isolate the separate movement components. The two components that are usually of most interest are the long-term movement and the seasonal fluctuation. If the time series shows a clear trend, one can interpolate successfully. Thus, for example, the trend in a time series of population statistics can be used to calculate values for the years in which no census was taken. If there is reason to believe that the factors that led to a particular economic time series will remain unchanged in the near future, this trend can be used to make forecasts in the short term. The seasonal fluctuation is usually used to draw up **seasonal indices** for the different months or terms in a year. Seasonal indices are calculated relative to the mean. This mean is indicated by 100; a January index of 118 in sales figures, for example, means that the January sales are 18% higher than the average sales.

1.5 The multiplicative model

When analysing time series the observed y values are usually accepted as being the product of the different movement components **actually** present in the series. In general the **multiplicative model** is expressed as

$$y = LSCI$$

Not all four movement components are necessarily present in a time series. Annual data, for instance, will show no seasonal fluctuation and consequently an acceptable model in this case would be $y = LCI$. A close look at the graphic representation of the annual rainfall figures (Figure 7.1) will reveal the absence of seasonal fluctuations and general trends. For this reason, $y = CI$ would be a useful model.

However, in this chapter, the **model for annual data** will be given as:

$$y = LCI$$

while the **model for monthly (or quarterly) data** will be expressed as

$$y = LSCI.$$

The multiplicative assumption enables us to use division to eliminate specific movement components from the time series. We shall consider two methods for determining L, the long-term trend. The first method, which is discussed in Section 2.1, entails fitting a curve by means of the least squares technique described in Chapter 6. The second method entails calculation of the **moving averages** and is discussed in Section 2.2. Suppose a curve $\hat{y} = f(x)$ is fitted to

the data. The value of L is then calculated as $\hat{y}$ and the value $y_i/\hat{y}_i$ is then the estimated value of the series at point x_i after eliminating the trend. In Section 3 a method for determining the seasonal fluctuation S is discussed. If the y values are divided by LS, we obtain the estimated CI values. These values are usually expressed as percentages and are used, *inter alia*, to determine the length of a cycle. This matter is discussed in greater detail in Section 4.

2 Long-term movement

2.1 The method of least squares

In most time series the long-term movement can be described by one of the curves represented graphically in Figure 7.4.

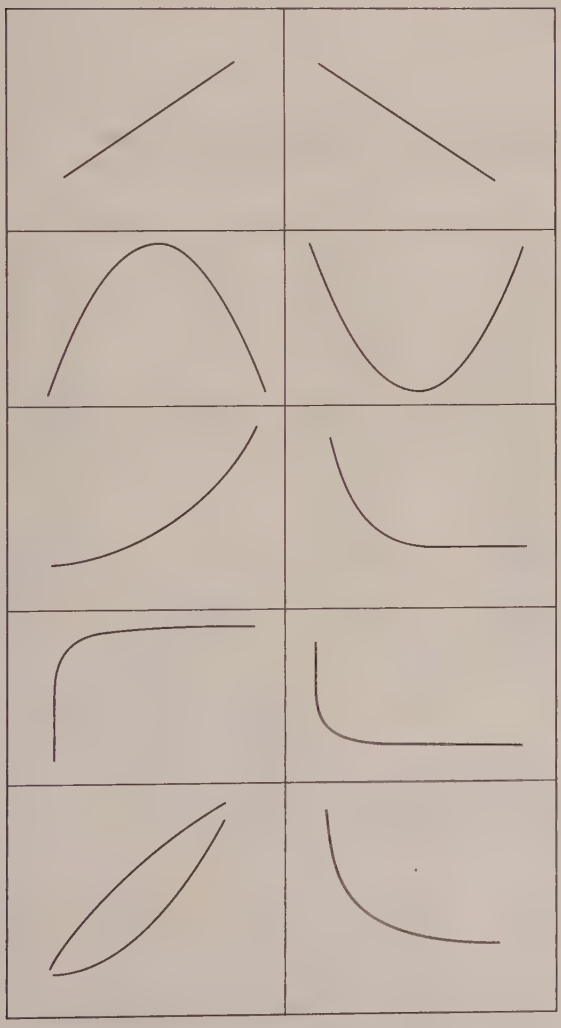


Figure 7.4 Illustration of possible trends

Chapter 6 included a full discussion on the procedures for fitting each of these curves to the data. Consequently we now confine ourselves to presenting two calculated examples. In the first example a least squares linear trend curve will be determined. The reader will be shown that the least squares trend curve is usually fitted after the seasonal fluctuation has been removed from the data. In the case of annual data the trend curve can consequently be determined directly.

EXAMPLE 7.1

Columns 1 and 3 of Table 7.1 reflect the value of coal sold annually between 1976 and 1988 (*S.A. Statistics, 1992*). The x values in column 2 were obtained by selecting 1982 as the origin. Note that $\sum x = 0$. In this case the normal equations for fitting a straight line are reduced to:

$$\sum y = na$$
$$\sum xy = b\sum x^2$$

Table 7.1 Value of coal sold (R million) for the period 1976 to 1988

Year (1)	x (2)	y (3)	xy (4)	x^2 (5)	$\hat{y}$ (6)
1976	−6	517.8	−3 106.8	36	−35.89
1977	−5	755.4	−3 777.0	25	440.02
1978	−4	874.4	−3 497.6	16	915.92
1979	−3	1 143.4	−3 430.2	9	1 391.83
1980	−2	1 471.3	−2 942.6	4	1 867.73
1981	−1	2 115.9	−2 115.9	1	2 343.63
1982	0	2 618.5	0.0	0	2 819.54
1983	1	2 620.1	2 620.1	1	3 295.44
1984	2	3 433.9	6 867.8	4	3 771.35
1985	3	5 038.1	15 114.3	9	4 247.25
1986	4	5 360.2	21 440.8	16	4 723.16
1987	5	4 788.3	23 941.5	25	5 199.06
1988	6	5 916.7	35 500.2	36	5 674.96
	0	36 654.0	86 614.6	182	

$$\bar{y} = \frac{36\,654}{13} = 2\,819.538$$

From Table 7.1 it follows that

$$a = \frac{\sum y}{n}$$
$$= 2\,819.538$$
$$b = \frac{\sum xy}{\sum x^2}$$
$$= 475.904$$

The equation of the least squares trend curve (see also Figure 7.5) is therefore:

$\hat{y} = 2\,819.538 + 475.9049x$
(Origin 1982, y units million rand, x units years).

An estimate of, for example, the 1992 trend is obtained if $x = 10$ is substituted in the above equation. It then follows that

$\hat{y} = 2\,819.538 + 475.9042(10)$
$= 7\,578.58$ million rand.

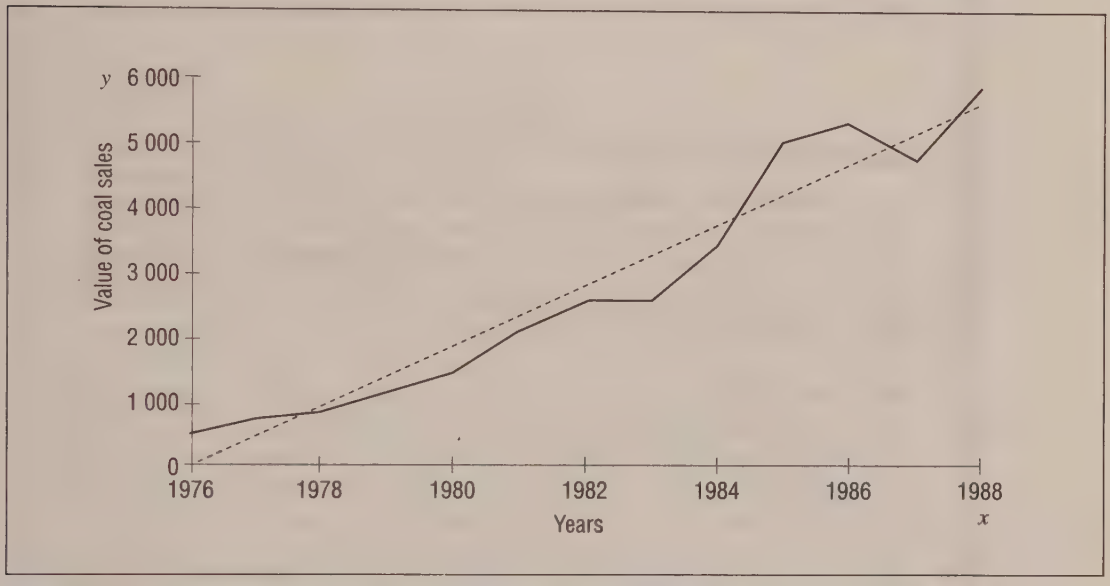


Figure 7.5 Least squares linear trend for the value of coal sold (million rand)

Notes

(a) When trend curves are fitted to time series data, one usually applies a transformation to the time scale values. In the case of an uneven number of years, the transformation

x_i = (ith year – middle year)

is usually used (see Example 7.1). This transformation has the advantage that $\sum x = 0$, which simplifies the normal equations considerably. If there is an even number of years, one of the following transformations is usually used:

(i) x_i = (ith year – first year)
(ii) x_i = 2(ith year – middle period)

Take the six years 1980, 1981, 1982, 1983, 1984, 1985 as an example. For the first transformation the x values are 0, 1, 2, 3, 4, 5 (origin 1980, units of years). For the second transformation, where $\sum x = 0$ also applies,

the x values are –5, –3, –1, 1, 3, 5 (origin 30 June 1983, units of half-years).

(b) When a hyperbolic or power function is fitted, the transformation must be such that none of the x values are zero. An appropriate transformation in this case is x_i = (ith year – first year) + 1. For the years 1980 to 1985 it therefore follows that $x_1 = 1$, $x_2 = 2$, ..., $x_6 = 6$.

(c) It is important to specify the origin and the x and y units.

EXAMPLE 7.2

As an illustration of the use of the transformation given in Note (b) above, we shall now fit the power curve $\hat{y} = ax^b$ to the sales of coal data (see Table 7.1). As mentioned in Chapter 6, this equation can be written in logarithmic form as $\hat{y}' = a' + bx'$, where y', a' and x' respectively indicate $\ln y$, $\ln a$ and $\ln x$. The least squares values of a' and b are given by

$$a' = \bar{y}' - b\bar{x}'$$

$$b = \frac{\sum x'y' - \frac{1}{n}\sum x' \sum y'}{\left\{\sum x'^2 - \frac{1}{n}\left(\sum x'\right)^2\right\}}$$

Table 7.2 Calculations for fitting a power curve to the data on coal in Table 7.1 ($x' = \ln x$, $y' = \ln y$)

Year	x	y	x'	y'	$x'y'$	x'^2	$\hat{y}$
1976	1	517.8	0.00000	6.24959	0.00000	0.00000	360.19
1977	2	755.4	0.69315	6.62725	4.59366	0.48045	736.47
1978	3	874.4	1.09861	6.77354	7.44149	1.20695	1 119.07
1979	4	1 143.4	1.38629	7.04176	9.76195	1.92181	1 505.83
1980	5	1 471.3	1.60944	7.29390	11.73908	2.59029	1 895.73
1981	6	2 115.9	1.79176	7.65724	13.71992	3.21040	2 288.13
1982	7	2 618.5	1.94591	7.87036	15.31501	3.78657	2 682.63
1983	8	2 620.1	2.07944	7.87097	16.36722	4.32408	3 078.94
1984	9	3 433.9	2.19722	8.14145	17.88860	4.82780	3 476.84
1985	10	5 038.1	2.30259	8.52478	19.62904	5.30190	3 876.15
1986	11	5 360.2	2.39790	8.58676	20.59014	5.74990	4 276.74
1987	12	4 788.3	2.48491	8.47393	21.05693	6.17476	4 678.49
1988	13	5 916.7	2.56495	8.68553	22.27796	6.57897	5 081.31
			22.55216	99.79706	180.38100	46.15387	

$\bar{x}' = \frac{22.55216}{13} = 1.73478$; $\bar{y}' = \frac{99.79706}{13} = 7.67670$

From the data in Table 7.2 it follows that

$$b = \frac{180.381 - \frac{1}{13}(22.55216)(99.79706)}{46.15387 - \frac{1}{13}(22.55216)^2}$$

$$= 1.03187$$

By substituting this value for b in the expression for a', we obtain

$$a' = 7.67670 - (1.03187)(1.73478)$$
$$= 5.88662$$

If a is obtained as the antilogarhithm of a' ($a = e^{a'}$), it follows that the power curve is given by

$$\hat{y} = 360.186x^{1.03187}$$

(Origin 1975, y units million rand, x units years).

The $\hat{y}$ values given in the last column of Table 7.2 are derived by substituting the x values concerned in the equation obtained.

2.2 The moving average method

A moving average can be regarded as an artificially created time series. The moving average series is obtained by replacing each observed y value with the averages of these specific values and the proximate y values. A three-year moving average for a series of values collected during 1989, 1990, 1991, 1992 and 1993 is obtained as follows: Add together the time series values corresponding to the first three years and write this total against the second year (1990). Subsequently the value of the first year (1989) is ignored and the total for the second to the fourth year is calculated. This total is written against the third year (1991). The process is then repeated until the last value in the series has been taken into account. The three-year moving averages are subsequently obtained by dividing each total by 3. Symbolically, we can express this as follows:

Year	y	Three-year moving average
1989	y_1	
1990	y_2	$\frac{y_1 + y_2 + y_3}{3}$
1991	y_3	$\frac{y_2 + y_3 + y_4}{3}$
1992	y_4	$\frac{y_3 + y_4 + y_5}{3}$
1993	y_5	

A five-year moving average is obtained in a similar manner. We begin by calculating the total of the first five values and by writing it against the third point, etc.

EXAMPLE 7.3

In the second column of Table 7.3 the annual production of condensed milk (in thousand tons, rounded off to the nearest hundred tons) is given with regard to the period 1970 to 1991 (see Appendix D, TR5.4). Columns 3 and 4 contain respectively the three-year moving total and three-year moving average, while columns 5 and 6 contain respectively the five-year moving total and the five-year moving average. Figure 7.6 is a graphic representation of the original time series and the five-year moving average. It is evident from this figure that in the five-year moving average, little is left of the irregular variation present in the original series. The gradual downward long-term trend is clearly visible.

Table 7.3 Calculation of a three-year and a five-year moving average for the annual production of condensed milk (in thousand tons) for 1970 to 1991

Year (1)	Production of condensed milk (2)	Three-year moving total (3)	Three-year moving average (4)	Five-year moving total (5)	Five-year moving average (6)
1970	36.4	–	–	–	–
1971	42.2	121.6	40.53	–	–
1972	43.0	130.5	43.50	211.8	42.36
1973	45.3	133.2	44.40	225.6	45.12
1974	44.9	140.4	46.80	237.1	47.42
1975	50.2	148.8	49.60	236.9	47.38
1976	53.7	146.7	48.90	234.8	46.96
1977	42.8	139.7	46.57	227.5	45.50
1978	43.2	123.6	41.20	218.9	43.78
1979	37.6	122.4	40.80	206.6	41.32
1980	41.6	120.6	40.20	200.7	40.14
1981	41.4	119.9	39.97	189.6	37.92
1982	36.9	110.4	36.80	182.7	36.54
1983	32.1	99.7	33.23	165.4	33.08
1984	30.7	87.1	29.03	147.0	29.40
1985	24.3	78.0	26.00	134.8	26.96
1986	23.0	72.0	24.00	126.4	25.28
1987	24.7	71.4	23.80	115.9	23.18
1988	23.7	68.6	22.87	114.3	22.86
1989	20.2	66.6	22.20	110.0	22.00
1990	22.7	61.6	20.53	–	–
1991	18.7	–	–	–	–

If there is an even pattern of variation in the data for which the period remains constant, this source of variation can be removed from the data by means of a moving average for the same period. Suppose, for example, that there is a seasonal fluctuation in a twelve-month data period. If the time series involved is observed on a monthly basis, a twelve-month moving average will elimate this seasonal component.

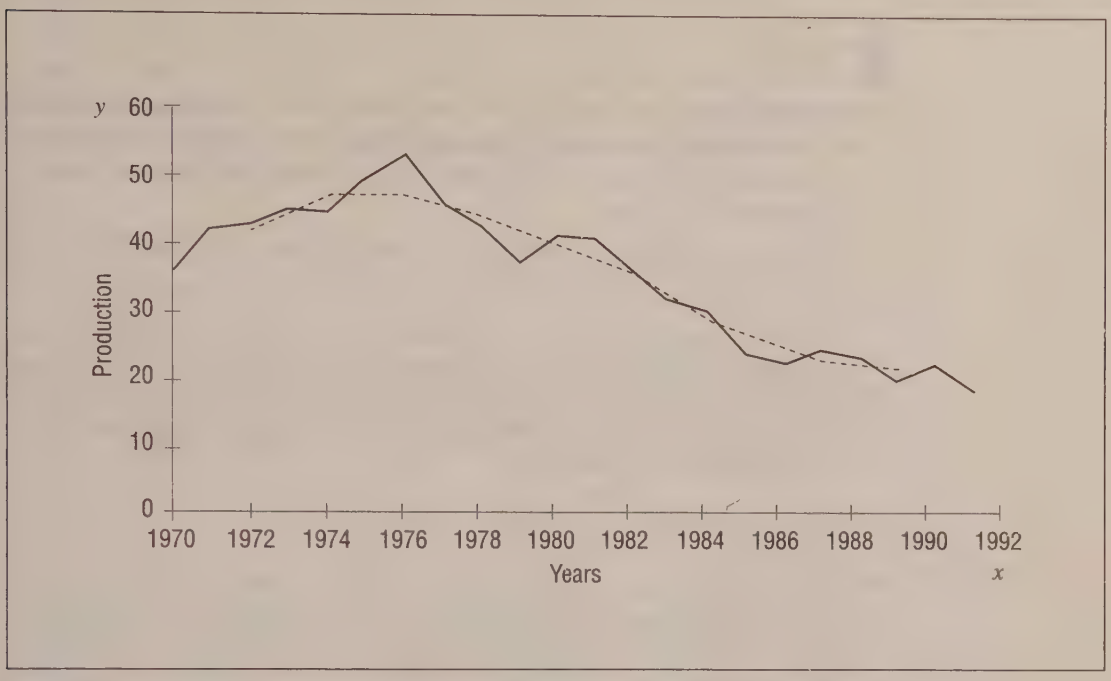


Figure 7.6 Annual production of condensed milk (in thousand tons). The dotted line indicates the five-year moving average

Since a moving average is constructed from the averages of previous, present and subsequent y values, irregular or incidental variations are, to a large extent, removed from the series.

There are two disadvantages attached to the use of a moving average for isolating the trend, namely:

(a) Periods of time are lost on both sides of the series. For a seven-year moving average, three years will, for instance, be lost at both the beginning and the end of the series.

(b) Since the moving average can neither be updated nor be expressed in terms of a specific mathematical equation, it is extremely difficult to use this technique to make objective forecasts.

In spite of these disadvantages, the moving average remains a useful technique for analysing time series. Firstly, a moving average enables us to form a visual image of the trend. In this way we can then fit an appropriate curve to the original data. Secondly, seasonal indices are based on moving averages. This concept is illustrated in the following section.

Notes

(a) Although we have computed a moving average by calculating a series of straight arithmetic averages, weighted averages can also be used. Consider, for instance, a five-year moving average. Each of the values in this series is an average of the form:

$$\bar{y} = \frac{y_{-2} + y_{-1} + y_0 + y_1 + y_2}{5}$$

Thus, for example, the notation y_{-2} denotes the observed y values two time units earlier than the present observation, y_0. Suppose the trend is poorly described by a straight line, but is well described by a polynomial of degree two or three. It can be demonstrated that a moving average based on a series of averages of the form

$$\bar{y}_g = \frac{-3y_{-2} + 12y_{-1} + 17y_0 + 12y_1 - 3y_2}{35}$$

will eliminate, from the data, a larger portion of the variation than a moving average based on equal weights. If the same conditions apply when a seven-year moving average needs to be calculated, the series of averages can be determined as follows:

$$\bar{y}_g = \frac{-2y_{-3} + 3y_{-2} + 6y_{-1} + 7y_0 + 6y_1 + 3y_2 - 2y_3}{21}$$

Note that the absolute values of the weights diminish on both sides of the middle value and the total of the weights = 1.

(b) If the moving average is based on an even number of observations, a so-called **centred moving average** is constructed by means of this series (see Example 7.4).

3 Seasonal fluctuation

In this section only monthly data will be analysed. However, the technique being discussed is also directly applicable to, for example, quarterly data. Calculating seasonal indices is the most important aspect of the analysis of seasonal fluctuation.

> A **seasonal index** contains twelve figures, one for each of the twelve months of the year (or years) and is used as a measure of activity for each month relative to the average activity over a year (or number of years).

There are thus two kinds of seasonal indices, namely a **specific seasonal index** quantifying the seasonal changes during a specific year and a **typical seasonal index** quantifying the seasonal changes over a number of years.

From the graphic representation of the specific seasonal indices one can establish whether the pattern of variation for the number of years over which the time series extends remains more or less stable. If this is not the case, there is no point in constructing a typical seasonal index.

3.1 Calculation of specific and typical seasonal indices

In Section 2 two properties of a moving average were given. These properties are repeated here:

If the seasonal fluctuation S has a pattern of variation that is repeated every twelve months, a twelve-month moving average will eliminate S from the data.

By taking averages of previous, present and subsequent y values, a moving average eliminates most of the irregular component I present in the data (see Figure 7.6 in this regard).

The combined effect of the above properties is that the twelve-month moving average eliminates most of the SI component from the data. If the model $y = LSCI$ is used as point of departure, it follows that the **twelve-month moving average is an estimation of the LC component**. A specific seasonal index will subsequently be calculated by dividing the original y value by the corresponding LC value, in other words:

$$\text{Specific seasonal index} \qquad SI = \frac{LCSI}{LC}$$

The next step is to eliminate the irregular variation from SI by calculating the averages for each month. In this way the typical seasonal index S is obtained.

EXAMPLE 7.4

The second column of Table 7.4 reflects the monthly wool exports (in million rand) for January 1988 to January 1993, as contained in Appendix D, time series TR9. Columns 3 and 4 are the twelve-month moving total and average. Note that the first moving average value has been entered between the months of June and July, the second moving average value between July and August, and so forth. The reason for this is the even span of the moving average, namely 12. To 'centre' these values against the months, the **centred moving average** is calculated (column 5). This is done by taking the average of the first and second values in column 4, and then the average of the second and third values, etc. The specific seasonal indices given in column 6 are obtained by dividing the y values in column 2 by the estimated LC values in column 5 and expressing this result as a percentage. Figure 7.7 is a graphic representation of the original time series together with the twelve-month centred moving average. In Figure 7.8 the specific seasonal indices for 1989, 1990 and 1991 are represented graphically.

Table 7.4 Calculation of specific seasonal indices for the monthly wool exports (million rand) for January 1988 to January 1993

	Date	Value of wool exports (million rand) y	12 month moving total	12-month moving average	Centred 12-month moving average LC	Specific seasonal index $SI = \frac{y}{LC} \times 100$
	(1)	(2)	(3)	(4)	(5)	(6)
1988	January	40.2				
	February	63.9				
	March	80.4				
	April	70.6				
	May	53.6				
	June	88.0				
	July	34.0	826.2	68.85	69.37	49.01
	August	52.6	838.6	69.88	72.40	72.66
	September	53.4	898.9	74.91	75.89	70.36
	October	105.8	922.5	76.88	77.42	136.66
	November	63.0	935.5	77.96	78.49	80.27
	December	120.7	948.2	79.02	78.71	153.34
1989	January	52.6	940.9	78.41	78.11	67.34
	February	124.2	933.8	77.82	76.80	161.72
	March	104.0	909.4	75.78	75.80	137.21
	April	83.6	909.7	75.81	74.56	112.13
	May	66.3	879.7	73.31	73.68	89.98
	June	80.7	888.7	74.06	71.92	112.21
	July	26.9	837.4	69.78	69.98	38.44
	August	28.2	842.1	70.18	69.13	40.80
	September	53.7	816.9	68.08	66.81	80.37
	October	75.8	786.6	65.55	65.73	115.33
	November	72.0	790.8	65.90	66.79	107.80
	December	69.4	812.1	67.68	66.65	104.12
1990	January	57.3	787.6	65.63	65.43	87.57
	February	99.0	782.8	65.23	65.54	151.06
	March	73.7	790.1	65.84	64.80	113.73
	April	87.8	765.2	63.77	62.24	141.06
			728.6	60.72		

Table 7.4 (Continued)

(1)		(2)	(3)	(4)	(5)	(6)
	May	87.6	709.3	59.11	59.91	146.21
	June	56.2	702.6	58.55	58.83	95.53
	July	22.1	723.7	60.31	59.43	37.19
	August	35.5	700.8	58.40	59.35	59.81
	September	28.8	696.1	58.01	58.20	49.48
	October	39.2	673.9	56.16	57.08	68.67
	November	52.7	665.0	55.42	55.79	94.47
	December	62.7	697.1	58.09	56.75	110.48
1991	January	78.4	716.0	59.67	58.88	133.15
	February	76.1	721.8	60.15	59.91	127.03
	March	69.0	733.7	61.14	60.65	113.78
	April	65.6	756.0	63.00	62.07	105.69
	May	78.7	784.6	65.38	64.19	122.60
	June	88.3	773.0	64.42	64.90	136.06
	July	41.0	744.0	62.00	63.21	64.86
	August	41.3	731.5	60.96	61.48	67.18
	September	40.7	730.5	60.88	60.92	66.81
	October	61.5	722.0	60.17	60.52	101.62
	November	81.3	688.4	57.37	58.77	138.34
	December	51.1	658.0	54.83	56.10	91.09
1992	January	49.4	656.7	54.73	54.78	90.18
	February	63.6	640.7	53.39	54.06	117.65
	March	68.0	649.7	54.14	53.77	126.47
	April	57.1	632.6	52.72	53.43	106.87
	May	45.1	609.5	50.79	51.75	87.14
	June	57.9	611.7	50.98	50.88	113.79
	July	39.7	598.4	49.87	50.42	78.74
	August	25.3				
	September	49.7				
	October	44.4				
	November	58.2				
	December	53.3				
1993	January	36.1				

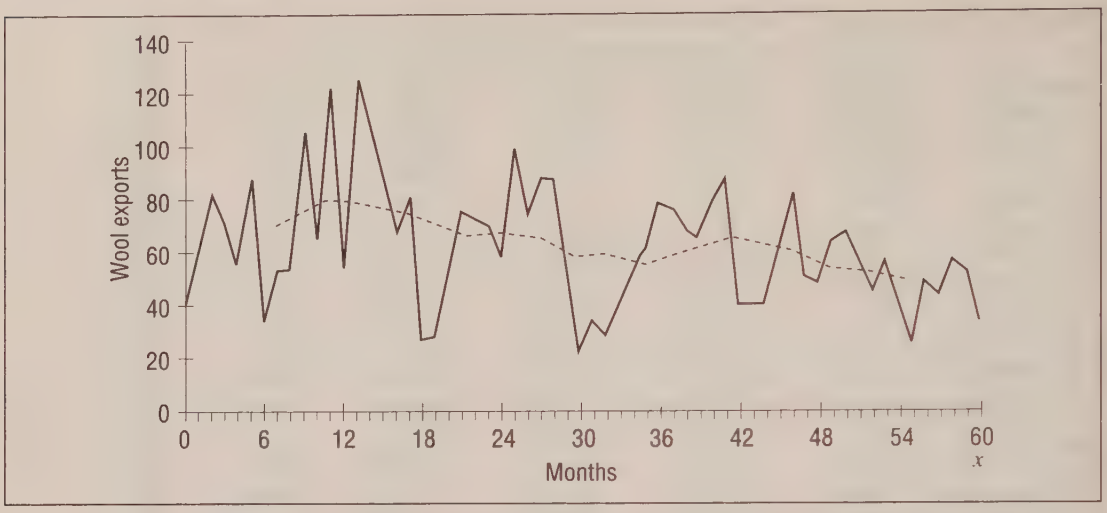


Figure 7.7 Graphic representation of the value of wool exports (million rand). The twelve-month centred moving average is indicated by the dotted line

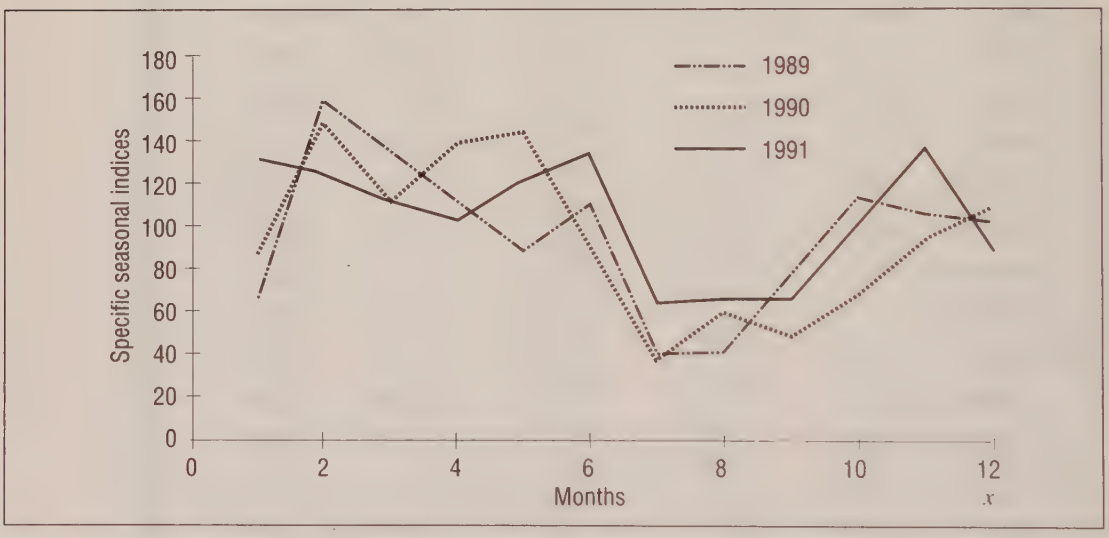


Figure 7.8 Specific seasonal indices for 1989, 1990 and 1991 for the value of wool exports

Since the pattern of variation is fairly stable, the typical seasonal index can now be calculated.

Table 7.5 is a summary of the specific seasonal indices. The second-last column of this table reflects the monthly averages. Note that the average monthly activity over the five years is 100.425. To calculate a seasonal index, each monthly average is multiplied by $\frac{100}{100.425}$. The typical seasonal index, given in the last column of Table 7.5, is thus calculated. Where there are extreme values under the specific seasonal indices, one usually calculates the median rather than the monthly averages.

Table 7.5 Seasonal indices for the value of wool exports, 1988 to 1992

Month	Specific seasonal index 1988	1989	1990	1991	1992	Monthly averages	Typical seasonal index
January		67.34	87.57	133.15	90.18	94.561	94.16
February		161.72	151.06	127.03	117.65	139.364	138.77
March		137.21	113.73	113.78	126.47	122.796	122.28
April		112.13	141.06	105.69	106.87	116.437	115.94
May		89.98	146.21	122.60	87.14	111.484	111.01
June		112.21	95.53	136.06	113.79	114.396	113.91
July	49.01	38.44	37.19	64.86	78.74	53.649	53.42
August	72.66	40.80	59.81	67.18		60.110	59.86
September	70.36	80.37	49.48	66.81		66.758	66.48
October	136.66	115.33	68.67	101.62		105.570	105.12
November	80.27	107.80	94.47	138.34		105.220	104.78
December	153.34	104.12	110.48	91.09		114.757	114.27
					Total	1 205.101	1 200.00
					Average	100.425	100.00

Figure 7.9 is a graphic representation of the typical seasonal index. The February activity appears to be approximately 39% higher than the average annual activity. As opposed to this, the August activity is approximately 40% lower than the average annual activity. In summation, one could say that wool exports are below average during January, July, August and September but above average during the remaining months.

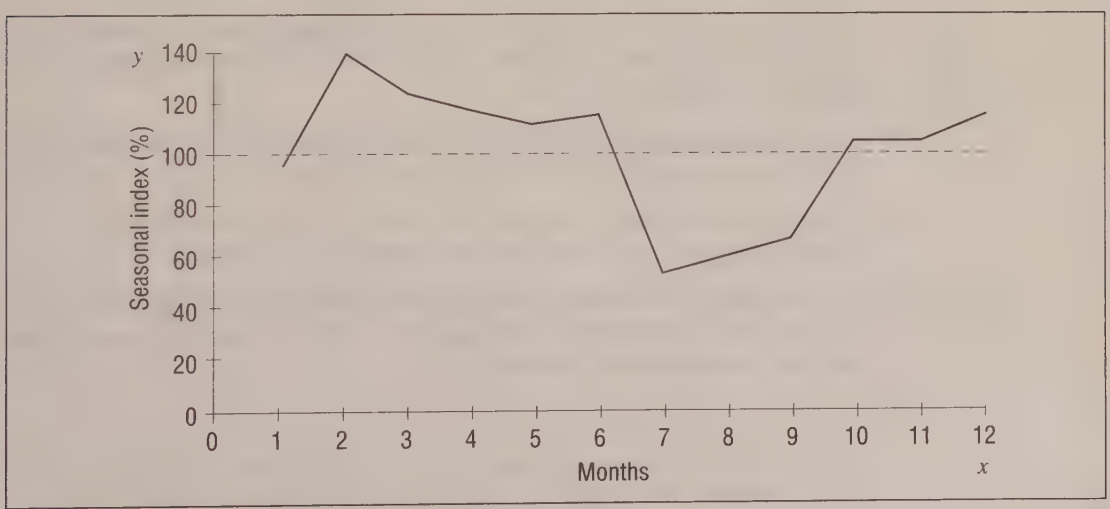


Figure 7.9 Typical seasonal index for the value of wool exports

Notes

(a) In the case of quarterly data a four-term centred moving index is calculated. The specific and typical quarterly indices are obtained in a way similar to the one used to calculate the monthly index.

(b) In addition to the original time series, official bodies sometimes also publish **seasonally adjusted time series**. The seasonally adjusted series $\{z\}$ of monthly wool sales given in column 4 of Table 7.6, Section 4 was obtained as follows: Each y value in column 2 of this table is divided by the typical seasonal index for that particular month $\left(\frac{LSCI}{S} = LCI\right)$. The typical seasonal index for wool exports, given in the last column of Table 7.5, was used for this purpose.

4 Cyclic movement

Cyclic movement seldom has a constant period of recurrence. The cycles in a particular series usually show a pattern of change that varies in intensity and length. As a result of this factor, the future behaviour of a cycle cannot be determined by extrapolation.

In the case of annual data ($y = LCI$), an estimation of the CI component can be obtained by eliminating L.

Since the real turning point in a cycle is often hidden in annual data, it is advisable to use monthly or quarterly data in determining the cyclic irregular component. For monthly data the procedure is basically the following:

(a) Calculate $LCI = \frac{LSCI}{S}$, with S the typical seasonal index, expressed as a ratio. The series thus obtained is called the **seasonally adjusted time series** (see Section 3, Note (b)).

(b) Fit a trend curve to the LCI series which was obtained in (a). The long-term trend (L) is estimated by $\hat{z} = f(x)$, where $z = LCI$.

(c) Eliminate L from the LCI series. This is done by dividing every LCI value by the corresponding $\hat{z}$ value, in other words $CI = \frac{LCI}{L}$, with $L = \hat{z}$.

(d) Express the CI values as percentages. A three-month or five-month moving average can subsequently be calculated in order to remove some of the irregular variation and so isolate C.

EXAMPLE 7.5

The second and fourth columns of Table 7.6 contain the monthly wool exports ($y = LCSI$) and the seasonally adjusted values ($z = LCI$) for time series TR9. (See Table 7.5 and Note (b) of Section 3.) In order to fit a linear trend curve to the z values, July 1990 is taken as the origin. The transformed time points (x) are given in the third column and the x units are months. Using the data from columns 3 and 4 we find that $\sum x = 0$, $\sum x^2 = 18\,910$, $\sum z = 3\,787{,}00$ and $\sum xz = -7\,637.00$. Since $\sum x = 0$, the normal equations are

$$\sum z = na \text{ and } \sum xz = b\sum x^2$$

From these equations it follows that

$$\hat{z} = 62.082 - 0.404x$$

The trend values given in column 5 are obtained by introducing the different values of x in the least squares equation. Note that only the first $\hat{z}$ value need be calculated. Thereafter $\hat{z}_2 = \hat{z}_1 + b$, $\hat{z}_3 = \hat{z}_2 + b$, etc. This method is, of course, only applicable if the x values are successive integers. Figure 7.10 is a graphic representation of the trend curve and the seasonally adjusted series.

In column 6 the cyclic irregular component is given as a percentage ($CI \times 100\%$). This series is derived by dividing the LCI values in column 4 by the L values from column 5. Column 7 is a three-month moving average based on the CI series and is regarded as an estimation of the business cycle. The cyclic irregular component and three-month moving average is represented graphically in Figure 7.11.

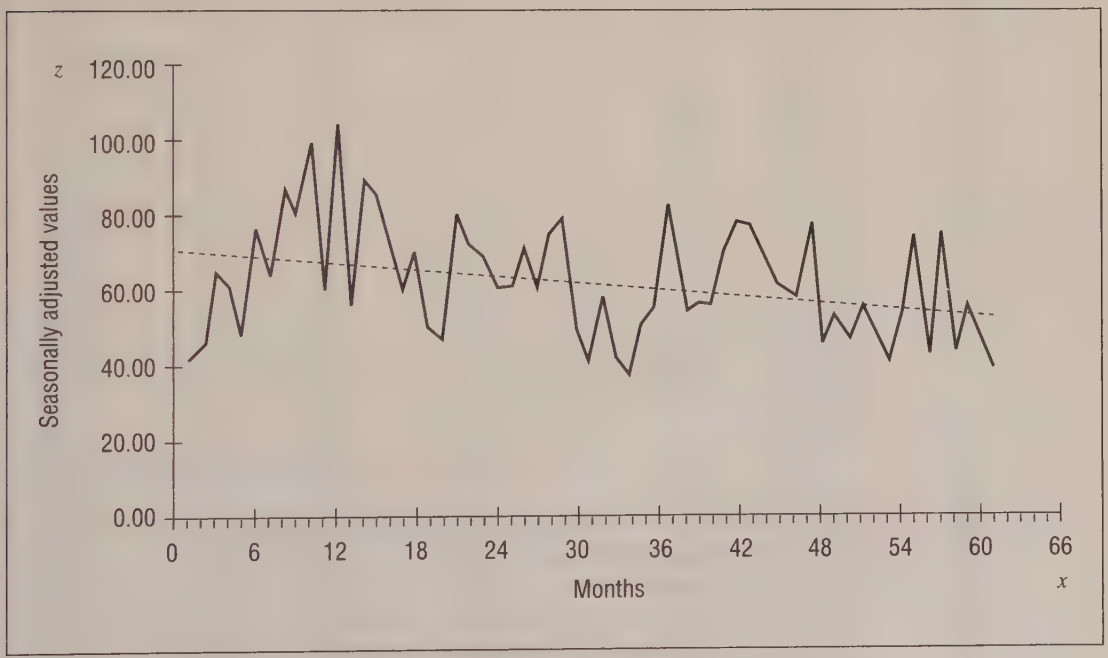


Figure 7.10 Graphic representation of the seasonally adjusted time series (LCI) of monthly wool exports (million rand). The straight line is the estimated trend curve

Table 7.6 Calculation of the business cycle for the monthly wool exports (million rand) for January 1988 to January 1993

	Date	Value of wool exports y	Trans-formed points x	Seasonally adjusted LCI values $LCI = z$	Long-term trend $L = \hat{z}$	Cyclic irregular component $CI \times 100$	Three-month moving average $C \times 100$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1988	January	40.2	–30	42.69	71.51	59.70	
	February	63.9	–29	46.05	71.20	64.67	72.37
	March	80.4	–28	65.75	70.89	92.75	81.23
	April	70.6	–27	60.89	70.58	86.28	82.58
	May	53.6	–26	48.28	70.27	68.72	88.48
	June	88.0	–25	77.25	69.95	110.44	90.18
	July	34.0	–24	63.64	69.64	91.39	109.53
	August	52.6	–23	87.88	69.33	126.76	111.51
	September	53.4	–22	80.33	69.02	116.39	129.88
	October	105.8	–21	100.64	68.70	146.49	116.93
	November	63.0	–20	60.13	68.39	87.92	129.85
	December	120.7	–19	105.63	68.08	155.15	108.50
1989	January	52.6	–18	55.86	67.77	82.43	123.42
	February	124.2	–17	89.50	67.46	132.68	113.93
	March	104.0	–16	85.05	67.14	126.67	122.41
	April	83.6	–15	72.10	66.83	107.89	108.12
	May	66.3	–14	59.72	66.52	89.78	101.56
	June	80.7	–13	70.84	66.21	107.01	91.07
	July	26.9	–12	50.35	65.89	76.42	85.09
	August	28.2	–11	47.11	65.58	71.84	90.67
	September	53.7	–10	80.78	65.27	123.77	102.20
	October	75.8	–9	72.11	64.96	111.00	113.69
	November	72.0	–8	68.72	64.65	106.30	103.90
	December	69.4	–7	60.73	64.33	94.40	98.59
1990	January	57.3	–6	60.85	64.02	95.05	100.48
	February	99.0	–5	71.34	63.71	111.98	100.70
	March	73.7	–4	60.27	63.40	95.07	109.03
	April	87.8	–3	75.73	63.08	120.04	113.61
	May	87.6	–2	78.91	62.77	125.71	108.25
	June	56.2	–1	49.34	62.46	78.99	90.42
	July	22.1	0	41.37	62.15	66.57	80.49
	August	35.5	1	59.31	61.84	95.92	77.63
	September	28.8	2	43.32	61.52	70.42	75.75
	October	39.2	3	37.29	61.21	60.92	71.31
	November	52.7	4	50.30	60.90	82.59	78.03
	December	62.7	5	54.87	60.59	90.56	103.77
1991	January	78.4	6	83.26	60.27	138.14	106.72
	February	76.1	7	54.84	59.96	91.45	108.06
	March	69.0	8	56.43	59.65	94.60	93.80
	April	65.6	9	56.58	59.34	95.35	103.35
	May	78.7	10	70.89	59.03	120.11	115.83
	June	88.3	11	77.52	58.71	132.03	127.85
	July	41.0	12	76.75	58.40	131.42	127.41
	August	41.3	13	69.00	58.09	118.78	118.72
	September	40.7	14	61.23	57.78	105.97	108.85
	October	61.5	15	58.50	57.46	101.81	114.52
	November	81.3	16	77.59	57.15	135.77	105.42
	December	51.1	17	44.72	56.84	78.67	102.42

Table 7.6 Continued

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1992	January	49.4	18	52.46	56.53	92.81	84.34
	February	63.6	19	45.83	56.21	81.53	91.27
	March	68.0	20	55.61	55.90	99.48	89.87
	April	57.1	21	49.25	55.59	88.59	87.19
	May	45.1	22	40.63	55.28	73.49	84.85
	June	57.9	23	50.83	54.97	92.47	100.65
	July	39.7	24	74.31	54.65	135.97	102.08
	August	25.3	25	42.27	54.34	77.78	117.38
	September	49.7	26	74.76	54.03	138.38	98.26
	October	44.4	27	42.24	53.72	78.63	107.01
	November	58.2	28	55.55	53.40	104.01	90.16
	December	53.3	29	46.64	53.09	87.85	88.17
1993	January	36.1	30	38.34	52.78	72.64	

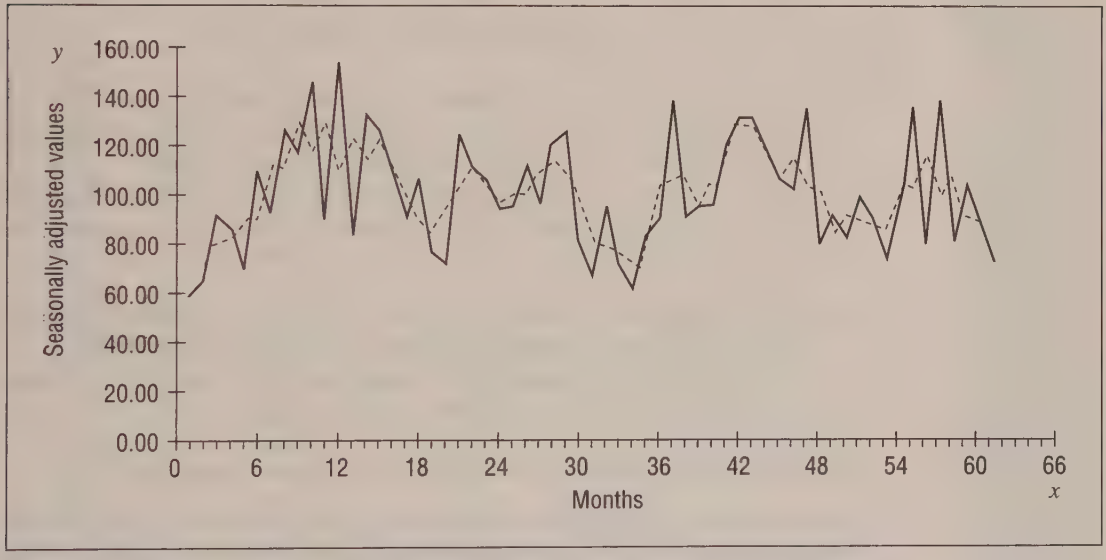


Figure 7.11 Graphic representation of the cyclic irregular component of TR9. The three-month moving average (C) is indicated by the dotted line

5 Concluding remarks

Long-term trends, seasonal movements and cyclic movements often occur in time series, although the measures according to which these components are determined are usually only approximations. Although previously perhaps not explicitly stated, reliable forecasts by means of the trend curve are based on the acceptance of independence between successive time series observations.

This however is not always the case and the analysis takes place by using models that assume dependency between successive y values. An example is the so-called Box Jenkins model. The analysis of this model is based on

advanced statistical theory which lies beyond the scope of this book. There are various program packages that can be used to carry out such analyses.

Despite the above remarks, the methods used in this chapter are very useful in analysing time series data. They can often be used to reliably predict the future behaviour of a time series.

Exercises

1 (a) Fit a least squares linear trend curve to time series TR4 of Appendix D for the period 1978 to 1990. Take 1984 as the origin and the x units in years.

(b) Use the first part of this exercise to estimate the trend for 1991 and 1992. Compare the estimated values with the observed 1991 and 1992 values.

2 Repeat Exercise 1 if a parabola is chosen as a trend curve.

3 (a) Fit a linear trend curve to time series TR5.4, namely the amount of condensed milk produced per annum (also see Table 7.3) for the period 1970 to 1991. Take 1970 as origin, x units as years.

(b) Draw a dotted line showing the estimated straight line and five-year moving average (Table 7.2, column (6)) on the same axis system for the period 1972 to 1989.

4 (a) Fit an exponential curve to time series TR7, namely electricity provided for domestic use for the period 1970 to 1986. Let 1971 be the origin.

(b) Use the estimated equation to predict the 1989 consumption. How does this differ from the observed consumption?

5 Repeat Exercise 4 if a power function is chosen as a trend curve.

6 Use the results in Exercise 1 to construct a *CI* series for the gross value of wheat production (TR4) and represent this series graphically. (Express the *CI* values as percentages.)

7 Calculate the five year moving average of time series TR5.1, namely the production of fresh milk for the period 1960 to 1991. Present the original time series and the moving average graphically on the same axis system.

8 Calculate the three year moving average for the amount of electricity provided for domestic use each year, of time series TR7 for the period 1967 to 1989 and represent this series graphically. Refer to Figure 7.4 and express your view on the type of curve that will best describe the long-term movement.

9 Construct the seven year and the eleven year moving average of time series TR3, the annual rainfall (mm) in Cape Town for the period 1960 to 1991.

Represent both series and the original time series graphically on the same axes.

10 Compile typical and specific quarterly indices for the time series TR13, namely the quarterly number of job opportunities in mining for the period January 1980 to December 1991.

11 Repeat Example 7.4 for the data of time series TR2.2 (end of month stock of wheat sorghum) for the period January 1984 to December 1992.

12 Repeat Example 7.4 for the data of time series TR2.1 and TR2.2, namely the end of month stock of wheat sorghum for the period January 1980 to December 1992. Note that the units of measurements have changed over this period (1 ton = 1 000 kg).

13 The table shows the production and value of silver from 1958 to 1991.

Year	Production (1 000 kg)	Value (million rand)
1958	56	1.1
1959	63	1.3
1960	69	1.5
1961	71	1.6
1962	79	2.0
1963	85	2.5
1964	91	2.7
1965	97	2.9
1966	97	2.9
1967	95	3.5
1968	104	5.2
1969	104	4.4
1970	110	4.6
1971	105	3.5
1972	102	4.1
1973	113	7.3
1974	84	8.2
1975	96	10.0
1976	88	10.7
1977	97	12.7
1978	97	14.7
1979	99	29.5
1980	191	81.5
1981	228	55.5
1982	240	53.7
1983	217	71.1
1984	217	66.2
1985	220	75.1
1986	237	73.9
1987	193	74.2
1988	200	79.5
1989	180	69.5
1990	175	57.3
1991	146	40.4

(a) Give a graphic representation of the production figures against years.
(b) Calculate the five-year moving average of the production figures and represent it graphically on the same axes that was used in (a).

14 Repeat Exercise 13 for the value of silver (column 3 of the table given for Exercise 13).

15 The following table represents the cost (rand) attached to the treatment of 1 ton of gold-bearing ore for the period 1964 to 1974:

Year:	1964	1965	1966	1967	1968	1969
Cost:	6.06	6.33	6.63	6.79	6.94	7.07
Year:	1970	1971	1972	1973	1974	
Cost:	7.34	7.88	8.79	10.51	13.18	

(a) Represent the time series graphically.
(b) Fit a straight line to the data using 1969 as origin.
(c) Calculate the expected treatment cost per ton for 1977.
(d) Estimate the 1971 *CI* value.

16 The following table shows the production of nickel for the period 1960 to 1974:

Year	1960	1961	1962	1963	1964
Production (tons)	3 376	2 915	4 876	2 073	4 175
Year	1965	1966	1967	1968	1969
Production (tons)	5 246	7 166	7 029	7 161	11 008
Year	1970	1971	1972	1973	1974
Production (tons)	11 557	12 757	11 656	19 426	22 100

(a) Give a graphic representation of both the time series and the three year moving average.
(b) Use (a) to decide on the form of a curve which should describe the trend well. Fit this curve to the data of 1968 to 1974 and estimate the production for 1976.

17 Which movement component(s) of a time series are important in the following problematic situations?
(a) A prediction of the demand for solar heating systems over the following five years is required for planning of the extension of factory facilities.
(b) A study by the department of Agriculture of the weather pattern in the Maize Triangle.
(c) The incidence of industrial unrest in the car manufacturing industry.
(d) The general growth of the steel industry over the last century.
(e) A tax rebate for the purchase of new houses leading to an increase in house sales.

18 The following data represents the sales (R100 000) of a beer producer for each quarter from 1990 to 1993:

Year	Quarterly sales I	II	III	IV
1990	24	19	25	38
1991	27	22	24	40
1992	31	23	23	41
1993	35	24	21	48

Do the following assignments (use one axis system for (a), (b) and (c)):

(a) Represent the time series graphically.
(b) Determine the least squares trend line and represent it graphically.
(c) Calculate centred four-quarterly moving averages and represent them graphically.
(d) Calculate the specific seasonal indices for 1983 and 1984 and the typical seasonal index and represent these indices graphically on the same axes.

19 The data below gives the average monthly bank debits (milliard rand) for 1976 to 1992 and is obtained by computing the annual averages of the monthly values (see TR8):

Year	Average bank debits
1976	29.29
1977	32.04
1978	38.62
1979	48.37
1980	68.17
1981	83.30
1982	97.68
1983	131.78
1984	162.65
1985	168.54
1986	195.84
1987	299.72
1988	276.73
1989	315.10
1990	406.50
1991	468.44
1992	616.08

(a) Represent the time series graphically.
(b) Fit an exponential curve to the data to represent the long-term movement.
(c) Use the fitted curve to predict the average monthly bank debits for 1992. How does this estimate compare with the real value? (see Appendix D, TR8).
(d) Estimate the average monthly value for 1995 and 2000.

20 The management of a holiday hotel would like to determine the seasonal pattern of the demand for accommodation so as to make provision for peak periods. The following table contains the quarterly occupation (number of guests):

Year	Quarterly sales			
	I	II	III	IV
1989	2 203	1 861	1 908	2 415
1990	2 343	1 921	1 986	2 514
1991	2 154	1 834	1 799	2 098
1992	2 025	1 837	1 967	2 304
1993	2 414	2 073	1 967	2 339

Calculate a typical seasonal index for the quarterly occupation of rooms.

21 The data below represents the electricity provision for distribution in South Africa (measured in Gigawatt hours) for the period January 1977 to December 1980 as published by the Central Statistical Service.

Month	1977	1978	1979	1980
January	6 178	6 449	7 059	7 683
February	5 814	6 080	6 701	7 603
March	6 446	6 710	7 526	8 156
April	6 215	6 803	7 291	7 910
May	6 908	7 260	7 866	8 761
June	6 930	7 543	7 985	8 821
July	7 401	7 846	8 591	9 205
August	7 215	7 542	8 392	8 919
September	6 676	7 249	7 839	8 399
October	6 766	7 333	7 930	8 445
November	6 470	7 082	7 711	8 180
December	6 333	6 915	7 519	7 885

Do the following assignments (use one axis system for (a) to (d)).

(a) Represent the time series graphically.
(b) Fit a least squares linear trend line to the year averages and represent this graphically.
(c) Calculate the five month moving average and represent it graphically.
(d) Calculate the centred twelve month moving average and represent it graphically.
(e) Calculate the specific seasonal indices for 1978 and 1979 and the typical seasonal index and represent them graphically on the same axis system.
(f) Estimate the provision of electricity for September 1982.
(g) Determine the cyclic irregular component.
(h) Isolate C by calculating a three-month moving average of the $CI \times 100\%$ value.

22 (a) Calculate a centred twelve month moving average over the period January 1989 to May 1993 for the values of the time series TR16 in Appendix D, namely the monthly index of the Johannesburg Stock Exchange for manganese shares, and represent it graphically.

(b) Compile specific seasonal indices and a typical seasonal index and represent them graphically on the same axis system.

23 Repeat Exercise 22 for the period January 1989 to May 1993 for the time series TR21, which is the monthly Johannesburg Stock Exchange index for all shares.

24 (a) Calculate the seasonally adjusted time series for the January 1990 to December 1992 values of the time series TR17, which is the monthly Johannesburg Stock Exchange index for coal mining shares (see Section 3, remark (b) and Section 4).
(b) Determine the cyclic irregular component and represent it graphically.

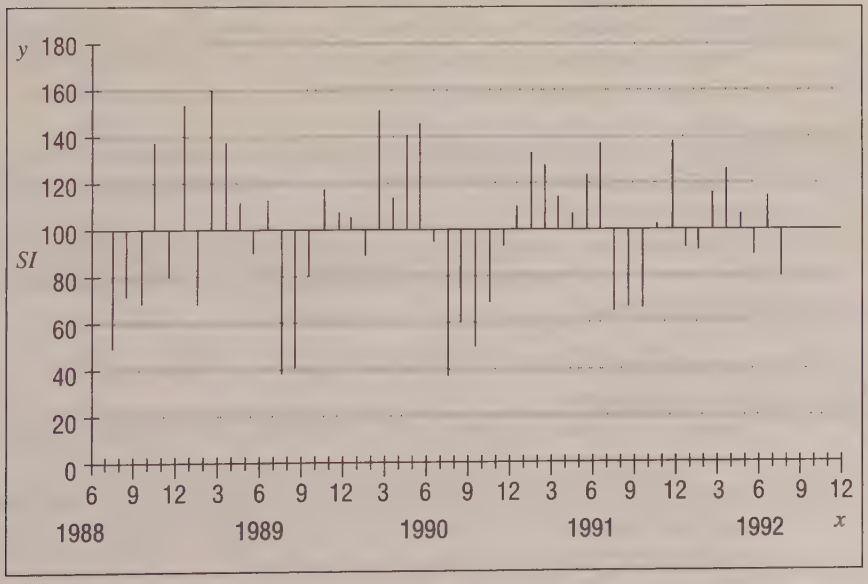
25 Repeat Exercise 24 for time series TR19, which is the monthly Johannesburg Stock Exchange index for financial shares.

26 Repeat Exercise 24 for time series TR18, which is the monthly Johannesburg Stock Exchange index for gold shares.

27 Repeat Exercise 23 for the consumer price index (all items) given in Chapter 8, Table 8.5.

28 Repeat Exercise 22 for the monthly inflation rate given in Chapter 8, Table 8.6.

29 Specific seasonal indices are represented alternatively by means of a needle diagram as illustrated below for the value of wool exports (TR6) (see Table 7.4, column (6)). Compile a needle diagram for the value of wool exports with regard to the irregular component (I) for the period July 1988 to July 1992 ($I = \frac{SI}{S}$, see Tables 7.4 and 7.5).



30 Represent the specific time series as well as the irregular movement graphically for time series TR9 namely the total value of wool exports over the period January 1990 to December 1992. Use the needle diagram representation as described in Exercise 29.

31 The following table shows the production (1 000 t), sales (1 000 t) and value of coal sold (R million) for the period 1974 to 1991.

Year	Production (1 000 t)	Sales (1 000 t)	Total sales (R million)
1974	66.1	64.7	200.0
1975	69.4	69.1	316.1
1976	76.5	75.8	517.8
1977	85.4	85.0	755.4
1978	90.4	85.7	874.4
1979	103.7	98.2	1 143.4
1980	116.1	111.2	1 471.3
1981	131.4	128.1	2 155.9
1982	144.2	137.1	2 618.6
1983	145.7	142.9	2 620.1
1984	163.0	159.6	2 433.9
1985	173.6	169.9	5 038.1
1986	176.8	172.4	5 360.2
1987	176.6	173.3	4 788.3
1988	181.4	185.9	5 916.7
1989	176.3	177.1	7 345.7
1990	174.8	184.1	8 149.6
1991	177.4	179.1	8 675.0

Use the coal production figures to answer the following questions:

(a) Give a graphic representation of the time series and the three-year moving average.

(b) Use (a) to decide on the shape of a curve that will adequately describe the trend. Fit this curve to the data for 1974 to 1991 and estimate the production for 1995.

32 Use the coal sales given in Exercise 31 to do the following assignments:

(a) Prepare a graphic representation of the time series and the three-year moving average.

(b) Use (a) to decide on the form of a curve that will adequately describe the trend. Fit this curve to the data for 1974 to 1991 and estimate the sales for 1995.

33 Use the value of coal sales (R million) given in Exercise 31 to do the following assignments:

(a) Prepare a graphic representation of the time series and the three-year moving average.

(b) Use (a) to decide on the shape of a curve that will adequately describe the trend. Fit this curve to the data for 1974 to 1991 and estimate the value for 1995.

34 The time series TR1 in Appendix D reflects the number of births for a certain section of the population between 1915 and 1990.
   (a) Prepare a graphic representation of the time series.
   (b) Fit a power curve to represent the long-term movement.
   (c) Use the fitted curve to estimate the number of births in 1995 for the specific section of the population.

35 Consider time series TR14 in Appendix D, which shows the number of motorcycles and scooters sold by garages between January 1987 and July 1992. Do the following assignments (use one axis system for (a) to (d)):
   (a) Prepare a graphic representation of the time series.
   (b) Fit a least squares linear trend to the year averages and present this in graphic form.
   (c) Calculate the five-month moving average and present this in graphic form.
   (d) Calculate the centered twelve-month moving average and present this in graphic form.
   (e) Calculate the specific seasonal indices for 1990 and 1991 and the typical seasonal index and present these indices on a single axis system.
   (f) Estimate sales for August 1992.
   (g) Determine the cyclic irregular component.
   (h) Isolate C by calculating a three-month moving average for the $CI \times 100\%$ values.

36 Consider time series TR11 in Appendix D which reflects the quantity of maize (1 000 tons) sold by producers between January 1988 and July 1992. Do the following assignments (use one axis system for (a) to (d)):
   (a) Prepare a graphic representation of the time series.
   (b) Fit a least squares linear trend to the year averages and present this in graphic form.
   (c) Calculate the five-month moving average and present this in graphic form.
   (d) Calculate the centered twelve-month moving average and present this in graphic form.
   (e) Calculate the specific seasonal indices for 1990 and 1991 and the typical seasonal index and present these on the same axis system.
   (f) Determine the cyclic irregular component.
   (g) Isolate C by calculating a three-month moving average for the $CI \times 100\%$ values.

37 Consider the number of road accidents given in Appendix D, TR10, for the period 1980 to 1992.
   (a) Prepare a graphic representation of the time series.
   (b) Fit a least squares linear trend to the yearly totals and present this in graphic form.
   (c) Calculate the three-year moving average and present it in graphic form.

38 Consider the annual production of factory butter given in Appendix D, TR5.2, for the period 1960 to 1991.
   (a) Prepare a graphic representation of the time series.
   (b) Fit a parabola to the annual totals and present this in graphic form.
   (c) Calculate the three-year moving average and present this in graphic form.
   (d) Estimate the annual production for 1992.

39 (a) Calculate the seasonally adjusted time series for the observations of January 1990 to December 1992 for time series TR20, which is the monthly Johannesburg Stock Exchange index for industrial shares (see Section 3, Remark (b) and Section 4).
   (b) Determine the cyclic irregular component and present this in graphic form.

40 Fit a least squares linear trend to the annual egg production (million dozen) for 1975 to 1991 (Appendix D, TR5.7) and estimate the egg production for 1995.

41 Fit a least squares linear trend to the annual production of factory-manufactured cheese (tons) for 1975 to 1991 (Appendix D, TR5.3) and make a forecast for production of cheese for 1995.

42 Fit a power curve to the annual production of full cream milk powder (tons) for 1975 to 1991 (Appendix D, TR5.5) and make a forecast for full cream milk powder production for 1995.

43 Fit a power curve to the annual production of skim milk powder (tons) for 1975 to 1991 (Appendix D, TR5.6) and make a forecast for skim milk powder production for 1995.

44 Time series TR12 in Appendix D gives the annual total value of gold sold between 1910 and 1991.
   (a) Prepare a graphic presentation of the value of gold sold against the years.
   (b) Calculate the five-year moving average of the sales figures and present this in graphic form on the axis used in section (a).

45 Consider the time series TR6 in Appendix D, which gives the monthly export value of gold (R million) for January 1988 to September 1992. Do the following assignments (use one axis system for (a) to (d)):
   (a) Prepare a graphic presentation of the time series.
   (b) Fit a least squares linear trend to the year averages and present this in graphic form.
   (c) Calculate the three-month moving average and present it in graphic form.
   (d) Calculate the centered six-month moving average and present this in graphic form.

(e) Calculate the specific seasonal indices for 1990 and 1991 as well as the typical seasonal index and present these indices in graphic form on the same axes.
(f) Determine the cyclic irregular component.
(g) Isolate C by calculating a three-month moving average for the $CI \times 100\%$ values.

46 Make a graphic presentation of the specific seasonal indices and the irregular movement for time series TR22 in Appendix D, which is the rainfall in the Katsedam catchment area during the period between January 1970 and September 1986, by using the needle diagram presentation set out in Exercise 29.

47 Consider Appendix A, Data Set 5, which reflects the weekly production of beer at a brewery.
(a) Present the beer production in graphic form as a time series with weeks on the x axis.
(b) Calculate the five-week moving average and present this in graphic form.
(c) Calculate the specific seasonal indices for 1992 and 1993 and present this in graphic form.
(d) Calculate the typical seasonal index and present this in graphic form.

48 Repeat Exercise 47 for the amount of electricity consumed by the brewery.

CHAPTER 8

Indices

1 Introduction

Change in virtually every stratum of society is a feature of our time. In addition to far-reaching political changes occurring not only in South Africa, but also in the world at large, demographic, economic and social changes affect the quality of everybody's life. It is important that the governments of countries, the industrialists and the providers of social services such as education, housing and health services measure these changes, determine the emerging trends and plan accordingly.

Price changes have a direct effect on the daily existence of every family. Every family has to constantly adapt its budget to make provision for price increases due to inflation. The continually rising cost of living leads to a demand for higher salaries which, if given, will increase production costs, which in turn can only be recovered by higher sales prices. Although this cycle does not benefit the economic development of any country, it does create a scenario that can be used to substantiate the calculation of various types of indices and to explain related concepts such as inflation and deflation. The aim of the calculation of indices (or index numbers) is to quantify changes in a standardized manner.

An **index** is a ratio that measures a relative change.

Although this is not essentially so, an index is usually expressed as a percentage. The following example illustrates the definition.

EXAMPLE 8.1

The information in Table 8.1 comprises building statistics obtained from Statistical News Release P 5041.2 of the Central Statistical Service.

The relative change in the value of building plans passed for 1990 and 1991 can now be represented meaningfully by the following indices:

Residential buildings	**Non-residential buildings**
$\frac{4\,932}{4\,212} \times 100\%$	$\frac{3\,304}{3\,169} \times 100\%$
$= 117.09\%$	$= 104.26\%$

Table 8.1 Value (million rand) of approved building plans and completed buildings (excluding additions and alterations) for the Republic of South Africa

	Building plans passed		Buildings completed	
	Jan–Dec 1990	Jan–Dec 1991	Jan–Dec 1990	Jan–Dec 1991
Residential buildings	4 212	4 932	2 681	3 057
Non-residential buildings	3 169	3 304	1 743	1 941

These indices are independent of the unit (million rand) used for the valuation of building plans. Between 1990 and 1991 there seems to have been an increase of 17.09% in the value of building plans passed for residential buildings. The increase in the value of passed plans for non-residential buildings was considerably smaller over the corresponding period, namely only 4.26%.

In respect of the value of completed buildings the following indices are obtained when 1991 is compared to 1990:

Residential buildings

$$\frac{3\,057}{2\,681} \times 100\%$$

$$= 114.02\%$$

Non-residential buildings

$$\frac{1\,941}{1\,743} \times 100\%$$

$$= 111.36\%$$

With regard to completed buildings, an increase is also noted in the values of residential buildings and non-residential buildings when 1991 is compared with 1990.

An index greater than 100% represents an increase, while an index smaller than 100% indicates a decrease. It is important to note the basis of comparison in the calculation of an index. Since the 1991 values in the above example were compared to the 1990 values (the 1990 values were used as denominators in the index ratio), 1990 is called the year of comparison or the **base year**. Although indices usually measure relative change over a period of time, they can also be used successfully with cross-section data. The number of residents in two cities or the salaries of two persons can thus be compared at any specific point in time. So also the value of passed residential building plans and the value of completed residential buildings in 1991 can be compared. This index is:

$$\frac{4\,932}{3\,057} \times 100\%$$

$$= 161.33\%$$

The provision of residences thus did not keep pace with the demand for residences, probably due to the deterioration within the building industry.

The quantification of change by means of indices is usually not as simple as the illustration in Example 8.1. An index is usually calculated as a measure of change in the prices, quantities or values of a group of items or of a sector as a whole. Individual changes in the different items have to be compounded to

obtain one representative index. The different ways in which this can be done, lead to a wide variety of indices, which will be discussed in Section 2 of this chapter. Problems that may develop around the choice of a base period, the selection of items to be used in the calculation of composite indices, and with the comparison of two series of indices, are dealt with in Section 3.

As price increases are a daily occurrence, it is important to have insight into compiling and interpretation of the **consumer price index**. This index is one of the most important economic indicators and plays an important role in decisions on, among other things, salary increases. The consumer price index, together with a few other special indices, is discussed in Section 4. In order to assess different types of indices objectively, certain mathematical characteristics have been defined and a good index must comply with these. These characteristics are explained in the last section of this chapter.

Because price changes are so important in modern society we shall work with price indices throughout this chapter, except where other indices are specifically mentioned. However, there is no problem in defining the corresponding quantity or value index with analogy to a specific type of price index figure.

2 Different types of indices

Before we discuss the different types of indices, we have to distinguish between simple and composite indices on the one hand, and unweighted and weighted indices on the other.

A **simple price index** is used to represent the price change of a single commodity, while a **composite index** is used to represent the price changes of more than one commodity.

Thus, for example, the indices calculated in Example 8.1 are called simple value indices. All the indices discussed in this section can be classified as composite indices.

When an **unweighted composite price index** is calculated, the price changes of all the commodities are regarded as equally important, while in a **weighted composite price index** different weights are allocated to the commodities according to the relative importance of each.

We shall now discuss these two types of composite price indices separately.

2.1 Unweighted composite price indices

The price of a specific commodity in the base period and the current period are indicated by p_0 and p_n. In order to explain the two most important unweighted

composite price indices, we consider the average prices of a number of important vegetables in Pretoria as observed in November 1988 and November 1992. The data in Table 8.2 was obtained from Statistical News Release P 0141.2 issued by the Central Statistical Service.

Table 8.2 Average prices (cent/kg) of seven types of vegetable in November 1988 and November 1992

Type of Vegetable	Prices (November 1988 = 100) November 1988 (p_0)	November 1992 (p_n)	p_n/p_0
Potatoes	107	267	2.495
Onions	114	176	1.544
Tomatoes	277	414	1.495
Green beans	296	742	2.507
Cabbage	38	218	5.737
Carrots	113	198	1.752
Pumpkins	122	182	1.492
Total	1 067	2 197	17.022

The notation (November 1988 = 100) indicates that November 1988 is the base month or month of reference, and that the prices of the current month (namely November 1992) are compared with those of the base month. The simplest way to summarize the price changes in one index, is to give the ratio of the totals of the November 1988 and the November 1992 prices, namely

$$P_1 = \frac{2\,197}{1\,067} \times 100$$
$$= 205.90$$

Since an index is always a percentage, the percentage sign will be omitted. The above index can generally be defined as follows:

The **simple composite price index** is given by

$$P_1 = \frac{\sum p_n}{\sum p_0} \times 100$$

If a simple price index for each individual type of vegetable is calculated, the arithmetic mean of these seven price indices can be used as an alternative index. This index is as follows:

$$P_2 = \frac{17.022}{7} \times 100$$
$$= 243.17$$

In general we can define this index as follows:

The **arithmetic mean of relative prices** is denoted by

$$P_2 = \frac{1}{k} \sum \left(\frac{p_n}{p_0} \right) \times 100$$

in which k = number of items.

The choice of selling units of the commodities has a substantial influence on the value of P_1. If the prices of potatoes are given as 'per pocket of 15 kg' instead of 'per kg', the value of P_1 changes to 231.14. However, P_2 is independent of the units in which the prices of the items are given. The arithmetic mean of relative prices, namely P_2, is therefore a more reliable composition of the price changes than the simple composite price index P_1.

2.2 Weighted composite price indices

It is important to remember that all commodities are not equally important. Appendix C contains the weights of the commodities that are used by the Central Statistical Service to determine the consumer price index. Under the heading *Foods* it can be seen that white bread, brown bread, beef, mutton, poultry, boerewors, fresh milk, cheese, potatoes, maize meal and eggs have the greatest weights. This means that these items are the most important foods in South Africa. Price changes in these important and essential foods are therefore of greater interest than price changes in unimportant foods such as sorghum meal, ham, fish paste, nuts, lettuce and vinegar. Naturally any composite index that does not take the relative importance of individual commodities into account does not reflect a true picture of the effect of price changes. Strictly speaking, the unweighted indices P_1 and P_2 are therefore only useful if they summarize the price changes of approximately equally important commodities. This necessitates a definition of weighted composite price indices.

The general formula for a **weighted composite price index** with weight series (w) is given by

$$P = \frac{\sum p_n w}{\sum p_0 w} \times 100$$

EXAMPLE 8.2

Supposing we need to calculate, for November 1992, the weighted composite price index for the seven types of vegetable in Table 8.2, using November 1988 as the base month. The weights of these items as indicated in Table 8.3 are found in Appendix C and although these weights are based on a 1990 consumer survey, we assume that they indicate the relative importance of each product in 1988 and in 1992.

Table 8.3 Prices (cent/kg) and weights of seven types of vegetable

Type of vegetable	Weight w	Price (November 1988 = 100) November 1988 (p_0)	November 1992 (p_n)	p_0w	p_nw
Potatoes	0.35	107	267	37.45	93.45
Onions	0.17	114	176	19.38	29.92
Tomatoes	0.26	277	414	72.02	107.64
Green beans	0.07	296	742	20.72	51.94
Cabbage	0.09	38	218	3.42	19.62
Carrots	0.09	113	198	10.17	17.82
Pumpkin	0.09	122	182	10.98	16.38
				174.14	336.77

The weighted composite price index for November 1992 with November 1988 as base is now calculated as

$$P = \frac{336.77}{174.14} \times 100$$
$$= 193.39$$

The general price index P, defined above, is often called the **fixed weight composite price index**. Special selections of the weight series $\{w\}$ lead to special types of composite price indices. If the quantities, $\{q_0\}$, of the items traded in the base period are used as weights for determining P, P reduces to the **Laspeyres price index**:

$$P_L = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100$$

If $\{q_n\}$ denotes the quantities traded in the current period and it is used as a weight series for P, we obtain the **Paasche price index**:

$$P_P = \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100$$

The merit of P_L lies in the fact that the weight series is fixed and can be used for as long as the base period is left unchanged. The denominator for P_L is therefore not recalculated if a new value is selected for n. As it is customary not to change the base period too often, it means that recent information on the quantities of the commodities included in the survey is not essential. On the other hand, P_P requires current information on the importance of each item concerned. Both the numerator and the denominator should therefore be recalculated if n changes.

If price changes are studied over a short term (one to five years), the difference between P_P and P_L is not marked, since consumer quantities of items do not change rapidly. For this reason the index $P = 193.39$ can be regarded as a relatively reliable indication of the increase in vegetable prices from November 1988 to November 1992 (see Example 8.2), although 1990 weights were used.

If the price of an item increases considerably from the base period to the survey period, the demand for the product usually drops. The base period weights therefore overestimate the actual weights of the commodities that undergo the biggest price increases. Commodities that undergo substantial price reductions, on the other hand, are subject to an increasing consumer demand, so that the base period weights in turn underestimate the relative importance of these items. The net result is that the quantity $\sum p_n q_n$ is expected to be large and P_L will reflect an overestimate of the actual price change. Similar arguments lead to the conclusion that P_P is expected to underestimate the actual price change. Unless drastic price and consumer changes occur in the period between the base period and the current period, the difference between P_L and P_P is small and both give good estimates of the price changes over that period. Each of these two indices is meaningful in the sense that a simple and an exact physical interpretation can be accorded to both.

If there is a relatively big difference between P_L and P_P, the interval between the two indices can be used as a zone of possible values. If, for example, $P_L = 130$ and $P_P = 120$, we can say that the price increase for the survey period was between 20% and 30%. If a single price index is required in such a case, a mean of the two indices is the logical choice. The following two price indices originated in this way. The arithmetic mean of P_L and P_P is known as the **Drobisch price index**:

$$P_D = \tfrac{1}{2}\left(P_L + P_P\right)$$

while the geometrical mean of P_L and P_P, namely

$$P_F = \sqrt{P_L \cdot P_P}$$

is known as the **Irving Fisher price index**.

2.3 The weighted arithmetic mean of relative prices

As the ratios $\{p_n/p_0\}$ are in themselves simple price indices and, as such, offer additional information on individual price changes, the ratios are also used in the definition of the **weighted arithmetic mean of relative prices**:

$$P_{wa} = \frac{\sum \left(p_n/p_0\right)w}{\sum w} \times 100$$

Although the quantities of commodities traded usually serve as weights to emphasize the relative importance of each item, the total amount of money spent on each item per unit of time can also be a meaningful weight series. Such weights are known as **value weights**. The money values of the items in the base period and in period n are given respectively by $\{p_0q_0\}$ and $\{p_nq_n\}$. If $w = p_0q_0$ is set in P_{wa} and simplified, it yields Laspeyres' index, namely

$$\frac{\sum (p_n/p_0)p_0q_0}{\sum p_0q_0} \times 100$$

$$= \frac{\sum p_nq_0}{\sum p_0q_0} \times 100$$

If $w = p_nq_n$ a further index is obtained, namely

$$P_3 = \frac{\sum (p_n/p_0)p_nq_n}{\sum p_nq_n} \times 100$$

The greatest disadvantage to the use of P_3 is that it is difficult to calculate and requires information regarding the quantities traded in year n.

EXAMPLE 8.3

Table 8.4 shows the April 1991 and April 1992 prices of seven car makes. Since each of these makes has a number of different models on the market, the median price was used in each case. The number of new motor vehicles of each of the seven makes registered in the first four months of 1991 and 1992 was used as the respective weight series and is also given in Table 8.4. The information with regard to the prices was obtained from *Car* magazine, while the registration figures were drawn from Statistical News Release P 7151.1 of the Central Statistical Service.

Table 8.4 Prices (R1 000) and registration figures (thousands) of seven car makes

Make	Prices		Number registered Jan – April	
	April 1991 p_0	April 1992 p_n	1991 q_0	1992 q_n
1. Audi	85.780	114.037	0.57	0.68
2. BMW	113.950	143.070	5.26	4.92
3. Ford	41.860	56.056	6.40	5.54
4. Mercedes Benz	160.285	207.284	2.78	2.29
5. Opel	31.210	40.986	7.54	5.97
6. Toyota	50.950	68.431	17.03	18.03
7. Volkswagen	38.170	49.830	13.43	12.29

The following table of calculations must be used to calculate the weighted composite price indices of Laspeyres, Paasche, Drobisch and Irving Fisher, and P_3:

	p_0	p_n	q_0	q_n	p_0q_0	p_0q_n	p_nq_0	p_nq_n	$\frac{p_n}{p_0}$	$\frac{p_n}{p_0} \times p_nq_n$
1.	85.780	114.037	0.57	0.68	48.895	58.330	65.001	77.545	1.329	103.057
2.	113.950	143.070	5.26	4.92	599.377	560.634	752.548	703.904	1.256	884.103
3.	41.860	56.056	6.40	5.54	267.904	231.904	358.758	310.550	1.339	415.826
4.	160.285	207.284	2.78	2.29	445.592	367.053	576.250	474.680	1.293	613.761
5.	31.210	40.986	7.54	5.97	235.323	186.324	309.034	244.686	1.313	321.273
6.	50.950	68.431	17.03	18.03	867.679	918.629	1 165.380	1 233.811	1.343	1 657.008
7.	38.170	49.830	13.43	12.29	512.623	469.109	669.217	612.410	1.305	799.195
					2 977.393	2 791.983	3 896.188	3 657.586	9.178	4 794.223

Laspeyres:

$$P_L = \frac{\sum p_nq_0}{\sum p_0q_0} \times 100$$
$$= \frac{3\,896.188}{2\,977.393} \times 100$$
$$= 130.86$$

Paasche:

$$P_P = \frac{\sum p_nq_n}{\sum p_0q_n} \times 100$$
$$= \frac{3\,657.586}{2\,791.983} \times 100$$
$$= 131.00$$

As these two indices differ from each other, it is meaningful to calculate the Drobisch and Irving Fisher indices as well:

Drobisch:

$$P_D = \tfrac{1}{2}\left(P_L + P_P\right)$$
$$= \tfrac{1}{2}(130.86 + 131.00)$$
$$= 130.93$$

Irving Fisher:

$$P_F = \sqrt{P_L . P_P}$$
$$= \sqrt{130.86 \times 131.00}$$
$$= 130.93$$

There thus seems to have been an increase of almost 31% in car prices in a single year from April 1991 to April 1992. This substantial price increase could perhaps have contributed to the fact that six of the seven car makes also showed a drop in the sales of new vehicles if the first four months of 1992 are compared to the corresponding four months in 1991.

The weighted and unweighted arithmetic means (with p_nq_n as value weights) of relative prices are, respectively:

$$P_2 = \tfrac{1}{k}\sum\left(\frac{p_n}{p_0}\right) \times 100$$
$$= \frac{9.178}{7} \times 100$$
$$= 131.11$$

$$P_3 = \frac{\sum\left(p_n/p_0\right)p_nq_n}{\sum p_nq_n} \times 100$$
$$= \frac{4\,794.223}{3\,657.587} \times 100$$
$$= 131.08$$

Notes

(a) Although the discussion thus far has been confined to the calculation of price indices, it is often important to determine the relative change, for instance, in sales of commodities or in the availability of electricity, or the relative change in the number of residents in a city. Indices that measure these changes, are called *quantity indices*. The quantities in the base year and the current year, namely q_0 and q_n, now fill the role of p_0 and p_n in all the above formulas. If a weighted composite quantity index needs to be calculated to indicate the change in the consumption of a group of items over, for instance, five years, the price series $\{p_0\}$ and $\{p_n\}$ can be used as the weights of the base year and of the current year. This is done because an increase in the consumption of an expensive item is far more important than the increase in the consumption of a cheap item. The Laspeyres quantity index will, for example, be as follows:

$$Q_L = \frac{\sum q_n p_0}{\sum q_0 p_0} \times 100$$

(b) The combined effect of quantity and price changes of a group of commodities is represented by the **simple composite value index**

$$W = \frac{\sum p_n q_n}{\sum p_0 q_0} \times 100$$

The denominator and numerator of W are respectively the total money value of the items in the base period and in period n.

(c) If the q_0 and q_n weights are known, the following two groups of weighted composite price indices also provide a reliable estimate of the relative price change of a group of commodities:

$$P_E = \frac{\sum p_n (q_0 + q_n)}{\sum p_0 (q_0 + q_n)} \times 100$$

$$P_4 = \frac{\sum p_n \sqrt{q_0 q_n}}{\sum p_0 \sqrt{q_0 q_n}} \times 100$$

The first index is calculated by using the arithmetic mean of q_0 and q_n, namely $\{q_0 + q_n)/2\}$as the weight series. This is known as the **Edgeworth price index**. The second index is obtained by using the geometrical mean of q_0 and q_n, namely $\left\{\sqrt{q_0 \cdot q_n}\right\}$ as the weight series. This index is less known and seldom used.

(d) The index P_2 is calculated as the arithmetic mean of the relative price series $\{p_n/p_0\}$. The **geometrical mean of relative prices**, P_5, is calculated in a similar manner from

$$\log P_5 = \tfrac{1}{k} \sum \log \left\{ \tfrac{p_n}{p_0} \times 100 \right\}.$$

3 Problems in determining and interpreting indices

3.1 The base period

If we compare the price of white bread in 1990 with that in 1993, 1990 is automatically regarded as the base year because it is the furthest in the past. Although this is not essential, a price index in year n is usually calculated with respect to a base year prior to that year (though not necessarily immediately prior). In all sectors of the economy the prices of items are usually listed monthly or, at least, annually so that new price indices can be calculated every month or year to give the most recent price changes. In the calculation of these indices the choice of a base period is often a problem. Unless the period directly prior to the survey period is used as the base period (see Section 5), the following should be kept in mind:

(a) The base period should not be too far in the past. A considerable number of commodities change in appearance or quality every year. The prices of a certain type of washing machine in 1970 and in 1990 will differ substantially, but the old and the new models will also differ. Many products disappear from the market while new products are introduced in the period between the base period and the current period. If the base period is too far in the past, such commodities will have to be ignored in the calculation of a price index. Such an index will then not be a true reflection of price changes.

(b) The base period should preferably be a period of reasonable economic stability. War years, periods of drought, strikes and serious recessions should therefore not be used as base periods. Institutions often purposely select the base period so that the calculated index actually gives a favourable impression. If a company compares its meagre profit in 1993 with a profit in 1980 that was, for certain reasons, the lowest annual profit for the past few decades, a high value for the profit index will not reflect the true state of affairs.

(c) If the prices and sales of commodities are available for a number of consecutive periods, it might be difficult or even meaningless to select a single base period. Instead, a reference price or sales quantity could be calculated for each item as the arithmetic mean of all the prices and sales quantities respectively. Relative changes in prices or sales of each item with respect to the arithmetic means then render valuable information.

When seasonal indices (see Chapter 7) are calculated, the mean monthly activity is used as the reference quantity with which each month's activity is compared.

3.2 Selection of items in composite indices

When a composite index is being determined, it is often impossible or unnecessary to include all items. Naturally it is an impossible task to include all commodities when determining, for instance, the consumer price index. If a representative sample of items is selected from the survey population and only these items are used in the calculation of the index, there is a time saving without undue loss of information. It is important that items included in the sample will not have undergone any essential change with regard to quality and appearance since the base period. In many cases items are selected from a population of items according to the method of judgmental sampling. The selection of items to be used in the calculation of the consumer price index will be fully discussed in Section 4.

3.3 Shifting the base period

The Bureau of Economic Research (BER) of the University of Stellenbosch publishes a monthly BER building costs index. It is commonly used to measure changes in building costs. The table below indicates the mean annual BER index values for the years 1984 to 1992.

1984	1985	1986	1987	1988	1989	1990	1991	1992	
287.8	299.2	322.2	369.9	453.8	549.1	630.4	702.5	723.4	(1975 = 100)

Although the indices are still published with 1975 as the base year, the base year can be shifted to, say, 1990, by multiplying each index by a constant. This constant should be such that the 1990 index is equal to 100. It is therefore clear that each index has to be multiplied by the value

$$\frac{100}{630.4} = 0.1586$$

For 1991 the BER index (1990 = 100) will thus be equal to

$$702.5 \times 0.1586 = 111.4$$

The BER index values for 1984 to 1992 with 1990 as the base year are then as follows:

1984	1985	1986	1987	1988	1989	1990	1991	1992	
45.6	47.5	51.1	58.7	72.0	87.1	100	111.4	114.7	(1990 = 100)

It is important to remember that the information in the above two tables is essentially the same. Although 1990 is the base year after the shift, the indices

are still being compared in terms of costs and weights obtained with 1975 as the base year. If the indices were to be calculated with 1990 as the base year, the index series obtained would generally differ somewhat from the above series. Despite this minor contradiction in index series, the shifting technique is often used to obtain 'comparable' indices. The ideal, however, is to adjust the weights of the items if the base year is adjusted. The shifting technique is also used to link two time series of the same index with different base periods. Such a link-up can, however, be made only if there is at least one point in time common to both series.

To illustrate: study the two BER index series below, namely for 1984 to 1987 with 1975 as the base year, and for 1987 to 1991 with 1985 as the base year.

1984	1985	1986	1987
287.5	299.2	322.2	369.9

(1975 = 100)

1987	1988	1989	1990	1991
123.5	151.6	183.4	210.6	234.6

(1985 = 100)

Since the two series have the 1987 index in common, they can be linked. If 1985 has to be the base year of the linked series, then it is clear that each of the indices in the first series should be multiplied by the constant

$$\frac{123.5}{369.9} = 0.3339$$

The linked series is then as follows:

1984	1985	1986	1987	1988	1989	1990	1991
96.0	100	107.6	123.5	151.6	183.4	210.6	234.6

(1985 = 100)

If each of the indices in the second series is multiplied by the constant

$$\frac{369.9}{123.5} = 2.9951$$

the linked series with 1975 as the base year is obtained.

4 Important indices

4.1 The consumer price index of South Africa (CPI)

This composite price index is a very important economic indicator. Salaried persons and pensioners must have insight into relative price changes in goods and services. In order to promote understanding of the practical problems experienced with the calculation of such an index, what follows is a relatively comprehensive description of the method for calculating current prices and

determining the weights of the selected items before the Central Statistical Service can calculate the CPI.

4.1.1 Base period

The base year is usually changed every five years. When the present edition of this book was written, the base year was 1990, but previously 1970 and 1975, 1980 and 1985 had also been used as base years.

4.1.2 Selecting items and determining weights

A comprehensive survey of the spending patterns of households is usually conducted every five years. The 1990 survey was conducted by the Human Sciences Research Council under contract to the Central Statistical Service. The survey was conducted in the following urban areas by means of a well-designed representative sample from all population groups:

Cape Peninsula	Pretoria/Verwoerdburg/Akasia
Port Elizabeth/Uitenhage	Witwatersrand
East London	Klerksdorp/Stilfontein/Orkney
Kimberley	Vaal Triangle
Pietermaritzburg	OFS Goldfields
Durban/Pinetown	Bloemfontein

The relative importance of each commodity was determined for each of the above twelve urban areas and for each of four groups (low income, middle income, high income and pensioners) in each area, i.e. for 48 strata, as a ratio:

$$\frac{p_0q_0}{\sum p_0q_0} \times 100$$

Here p_0q_0 represents the total amount per time unit that the households in the stratum concerned spend on that item. $\sum p_0q_0$ indicates the total expenditure of all the households in the stratum. For each group the weighted arithmetic mean of the item weights for the twelve urban areas was calculated. A 'total' weight series (base year weights) for all the households was calculated as a weighted arithmetic mean across the four groups. The items were then arranged in order of importance according to the latter weight series and the most important (almost 600) were divided into seventeen main groups, namely:

1. Food
2. Non-alcoholic beverages
3. Alcoholic beverages
4. Cigarettes, cigars and tobacco
5. Clothing and footwear
6. Housing
7. Fuel and power
8. Furniture and equipment
9. Household operation
10. Medical care and health expenses

11. Transport
12. Communication
13. Recreation and entertainment
14. Reading matter
15. Education
16. Personal care
17. Other goods and services

Household expenditure on these items represents approximately 80% of the total expenditure on goods and services. The weights of these items appear in Appendix C. Only the weights of the three income groups and pensioners and the 'total' weights are given in this appendix. Note that for the calculation of the consumer price index, income tax expenses, contributions to pension and provident funds and capital expenditure (such as capital redemption on house loans) is not taken into account.

4.1.3 Listing of prices

In order to obtain the prices of the selected items in the base year and subsequent periods, a representative sample of about 3 500 institutions was selected. With regard to goods supermarkets, department stores, general retailers and specialized retailers – like men's and ladies' outfitters and butcheries – were included in the sample. Also included were service institutions such as hairdressers, municipalities, control boards and film theatres. All these institutions send price quotations to the Central Statistical Service on a monthly basis (first week of every month). The Service prescribes each commodity or service involved with regard to quality and selling quantity as accurately as possible so that price quotations are comparable.

4.1.4 Calculation of index

A weighted arithmetic mean of relative prices with a weight series $\{w\}$, namely

$$P_v = \frac{\sum (p_n \; p_0) w}{\sum w} \times 100$$

is calculated every month for each of the above strata by the Central Statistical Service. When these indices are calculated, the items in Appendix C are used. As

$$w = \frac{p_0 q_0}{\sum p_0 q_0}$$

this index reduces to P_L, as shown in Section 2. A joint index for all the people in the country is calculated monthly and this index is known as the consumer price index. A price index of food alone is also calculated monthly. Table 8.5 contains both these indices for 1988 to 1993.

Table 8.5 Consumer price index for South Africa 1988 – 1993

	Consumer prices – All items (1990 = 100)					
	1988	1989	1990	1991	1992	1993
January	72.2	81.8	94.2	107.7	125.1	137.2
February	72.2	82.5	94.8	109.0	126.2	137.6
March	73.7	83.9	96.4	110.0	127.3	139.6
April	74.5	84.9	97.3	111.5	128.9	143.1
May	75.1	86.3	98.3	113.2	129.9	143.7
June	75.4	87.2	99.0	114.0	131.2	144.3
July	76.2	88.0	99.7	115.5	132.4	145.5
August	77.2	89.2	101.3	117.1	133.8	146.2
September	78.3	90.0	102.9	118.7	134.7	146.9
October	79.1	90.8	103.6	121.0	135.2	147.9
November	79.8	91.8	105.8	122.2	135.7	
December	80.6	92.9	106.5	123.8	135.7	
	Consumer prices – Food only (1990 = 100)					
	1988	1989	1990	1991	1992	1993
January	72.8	82.5	94.8	109.4	138.1	158.7
February	73.5	82.1	95.0	109.9	139.6	159.2
March	74.9	83.0	95.5	110.7	142.7	158.3
April	77.1	83.1	97.0	113.2	145.3	157.7
May	77.5	84.7	98.0	115.1	147.1	158.0
June	76.9	85.5	98.5	115.1	148.8	158.1
July	77.9	85.6	99.6	116.8	152.3	158.9
August	78.6	86.6	101.9	120.2	154.4	159.6
September	79.6	87.7	102.9	123.2	157.2	160.9
October	80.9	89.2	103.5	130.1	157.7	162.2
November	81.0	90.8	105.7	134.1	157.2	
December	81.2	93.1	107.4	137.8	157.1	

Notes

(a) These indices can be corrected by eliminating seasonal effects. However, these adjustments have only a minor effect.

(b) The consumer price index cannot be used as a measure of the difference in the cost of living between two urban areas. Although the change in the price of petrol in Johannesburg and in Port Elizabeth is always more or less the same, petrol remains cheaper in Port Elizabeth than in Johannesburg. When a substantial price increase occurs, the consumer will also switch, temporarily or permanently, to a replacement product that is cheaper to keep his cost of living as low as possible. The consumer price index is based on weights of the base year and therefore cannot take these changes in consumer preferences into account. This causes the consumer price index to overestimate the increase in the cost of living.

(c) The first retail price information in South Africa appeared in a price index calculated in 1895, entitled: *Price Index: Cost of Living (food, fuel, lights and rentals)*. This index was published regularly until 1910, after which a retail price index covering all items was used until 1937. National consumption was used as the point of departure for determining the weights. The first survey of family expenditure was conducted in 1936 among white urbanites to obtain a more reliable reflection of the relative importance of each commodity. Since 1958 the index measuring price changes in retail goods has been known as the consumer price index.

(d) The **inflation rate** in the ith month is determined as follows by means of the consumer price index P_v:

$$\text{Inflation rate } (i) = \left\{ \frac{P_v(i)}{P_v(i-12)} - 1 \right\} \times 100\%$$

Thus, for example:

$$\text{Inflation rate (Dec 1992)} = \left\{ \frac{135.7}{123.8} - 1 \right\} \times 100\%$$
$$= 9.61\%$$
$$\text{Inflation rate (Feb 1992)} = \left\{ \frac{109.0}{94.8} - 1 \right\} \times 100\%$$
$$= 14.98\%$$

Table 8.6 reflects the inflation rates for the years 1991 to 1993 on a monthly basis. This table shows a drastic drop in the inflation rate, especially from October 1992. The introduction of VAT at a rate of 10% on 1 October 1991, compared to the GST rate of 13% in the preceding period, probably contributed to this drop in the inflation rate.

Table 8.6 Inflation rate in South Africa 1991–1993

	1991	1992	1993
January	14.33	16.16	9.67
February	14.98	15.78	9.03
March	14.11	15.73	9.66
April	14.59	15.61	11.02
May	15.16	14.75	10.62
June	15.15	15.09	9.98
July	15.85	14.63	9.89
August	15.60	14.26	9.27
September	15.35	13.48	9.06
October	16.80	11.74	9.39
November	15.50	11.05	
December	16.24	9.61	

4.2 The Johannesburg Stock Exchange indices

One of the best-known indices reflecting price changes in shares on the Johannesburg Stock Exchange is the JSE overall index. Values for this index have been calculated and published since 1960. It is a Laspeyres index based on the share prices of approximately 130 different companies. The companies included are those that represent more than 80% of the market capitalization or value of the exchange.

Market capitalization is calculated by multiplying the total number of shares a company has issued by the price per share. The results are used effectively as weights in the calculation of this index. Similar individual indices for 47 different sectors are calculated daily and the closing value is published in the press. The price indices for gold shares, industrial shares (TR20, Appendix D) and financial shares (TR19, Appendix D) are particularly important and are regularly broadcast in financial news on, for example, television. Time Series TR16, TR17 and TR18 in Appendix D indicate the month-end values of the JSE index series for manganese, coal and gold.

4.3 The Dow Jones industrial shares index

This index, known as the Dow Jones industrial average, is a measure of price changes on the New York stock exchange. It was calculated for the first time in 1897 and was initially based on twelve different shares. At present it is calculated as the mean price of one share each, of thirty companies. These shares are not a random sample from the approximately 1 200 shares listed on the market, but were deliberately selected to represent the total spectrum of shares on the market. The Dow Jones average is calculated every day on the basis of the closing prices of the thirty shares. Although this index has been criticized considerably during the past few decades and had to compete with alternative price indices, it remains one of the best-known and most important international economic indicators.

4.4 The intelligence quotient (IQ)

Different tests are used to measure the intelligence of individuals. The intelligence quotient of, for example, a twelve-year-old child is determined by the ratio of this child's test score to the mean test score of all twelve-year-olds. The intelligence quotient is therefore an index that has been standardized so that the mean of all individuals of a certain age group is 100. An individual with an IQ of, for instance, 120, has an above-average intelligence, and one with an IQ of 90 has a below-average intelligence. The intelligence quotients of children of the same age group is usually distributed symmetrically around the mean of 100. The distribution is bell shaped and is adequately approximated by the so-called normal distribution. If the standard deviation of this distribution is known (it is often assumed to be 15), the quantiles can be found and used to interpret individual scores.

4.5 Other indices

A few important indices that measure change in divergent situations and are published regularly by the Central Statistical Service in its News Releases, and in other summary publications such as *South African Statistics* and *South African Labour Statistics*, are the following:

Production price index
Construction price index
Agricultural price index
Index of physical volume of factory production
Indices of volumes of import and export according to SITC sections and divisions.

5 Index chains

In many situations it is important to measure relative price and quantity changes only with regard to the directly preceding period. For example, a trader needs to know at the end of each month how the turnover of the past month compares with that of the previous month, while a car enthusiast wants to compare the prices of this year's models with those of the previous year's models. Suppose a price index is calculated in ten consecutive years with the preceding year as the base year in every case. The row of indices then forms a chain with links of consecutive indices. Symbolically such a chain of indices can be represented by $\{P_{i-1,i}\}$. The four indices for the years 1990 to 1993, for instance, form the chain

$$\{P_{89,90};\ P_{90,91};\ P_{91,92};\ P_{92,93}\}$$

Any one of the different types of indices defined in Section 2 can be used to obtain an index chain. For example, if we use the simple composite price index chain for the years 1991 to 1993, we obtain the following interesting result:

$$\begin{aligned} P_{90,91} \cdot P_{91,92} \cdot P_{92,93} &= \frac{\sum p_{91}}{\sum p_{90}} \cdot \frac{\sum p_{92}}{\sum p_{91}} \cdot \frac{\sum p_{93}}{\sum p_{92}} \\ &= \frac{\sum p_{93}}{\sum p_{90}} \\ &= P_{90,93} \end{aligned}$$

where the indices are expressed as ratios and not as percentages. The product of the elements of a simple composite price index chain for the years n_1 to n_2 therefore gives the price index of year n_2 with year n_1 as the base year. The same result is found in a weighted composite price index chain with fixed weights $\{w\}$, because

$$\begin{aligned} P_{90,91} \cdot P_{91,92} \cdot P_{92,93} &= \frac{\sum p_{91} \cdot w}{\sum p_{90} \cdot w} \cdot \frac{\sum p_{92} \cdot w}{\sum p_{91} \cdot w} \cdot \frac{\sum p_{93} \cdot w}{\sum p_{92} \cdot w} \\ &= \frac{\sum p_{93} \cdot w}{\sum p_{90} \cdot w} \end{aligned}$$

The reader can discern that this product rule does not apply to, for example, the indices of Laspeyres and Paasche. However the product of the 'links' still provides useful information in regard to price change over the whole period.

An advantage of the calculation of index chains is that new items that appear on the market can be included from year to year, while obsolete items can be omitted without the whole chain of indices having to be recalculated. The weights of the items concerned can also be changed from year to year if necessary. Naturally the multiplicative rule illustrated above for the simple composite index and the weighted composite index does not apply in the case of such adjustments of the included items and weights.

6 Deflation

The consumer price index can be successfully used to compare the present purchasing power of a currency with the mean purchasing power of the currency at a previous point in time. According to Table 8.5, the consumer price index for February 1993 equals 137.6. This means that in February 1993, R137.60 was paid for the same merchandise and services that were available for R100 on average in 1990. In other words: Compared with 1990, the buying power of R1,00 in February 1993 was

$$\text{R}\frac{100}{137.60} = \text{R}0.73$$

It is interesting to compare the current purchasing power of the rand with its purchasing power in, say, 1910. The following is an extract from a table in *RSA Statistics in Brief 1992*, the 1992 edition of a publication that is distributed annually by the Central Statistical Service, and indicates the diminishing purchasing power of 100c in specific years since 1910:

1910	1930	1950	1970	1975	1980	1985	1988	1989	1990	1991
100	72	47	25	16	9	5	3	3	2	2

It is clear that a salary that remains constant from one year to the next actually, in **real terms**, decreases from year to year. For example, if a person has received a fixed annual salary of R30 000 since 1990, his actual income in February 1993 was only R21 900 per annum if compared to the 1990 income.

Deflation is the process according to which an amount is corrected for relative price changes in merchandise and services.

The annual or biennial adjustment of salaries and wages in all sectors has become customary. To see whether the salary increases given to a person in a certain post over the past few years have kept pace with overall price increases, the salary for each year can be deflated to a base-year salary by using the consumer price index.

EXAMPLE 8.4

Table 8.7 reflects the annual salaries linked to three posts, namely a CS educator (standard scale, male, qualification category D), a programmer and an engineer, all in the public sector for the period 1988 to 1993. All salaries indicate the penultimate notch on the salary scale concerned.

Table 8.7 Annual salary (R) and deflated salary (R) of the three posts in the public sector for 1988 to 1993. The salaries as well as the CPI are as at 1 April of each year

1990 = 100		Annual salary (excluding fringe benefits) at 1 April		
Year	CPI	CS Educator	Programmer	Engineer
1988	74.5	23 991 (32 203)	19 857 (26 654)	40 527 (54 399)
1989	84.9	25 671 (30 237)	22 857 (26 922)	46 617 (54 908)
1990	97.3	36 387 (37 397)	30 636 (31 486)	51 279 (52 702)
1991	111.5	36 387 (32 634)	30 636 (27 476)	51 279 (45 990)
1992	128.9	41 163 (31 934)	32 898 (25 522)	56 655 (43 953)
1993	143.1	45 240 (31 614)	35 949 (25 122)	61 914 (43 266)

The deflated salary is obtained by dividing the salary by the corresponding consumer price index and multiplying the quotient by 100. Thus, for example, we find that, for 1991, the deflated salary of a CS educator is

$$\text{R}\left(\frac{36\,387}{111.50}\right) \times 100 = \text{R}\,32\,634$$

The amounts in brackets indicate the respective deflated salaries, that is, the salaries in 1990 rands. Table 8.7 indicates that the salaries of the CS educator and the programmer remained relatively constant over six years, while the salary of the engineer diminished.

EXAMPLE 8.5

Table 8.8 indicates the gross domestic product (GDP) for the RSA at current prices for 1980 to 1990, as well as the total population of the RSA and the CPI for the corresponding years. The penultimate column provides the per capita GDP for the various years and is an indication of the welfare in the RSA during these years. In order to view these amounts in perspective, the purchasing power is calculated in the final column in terms of 1990 rands. The drop in the values in the final column from 1985 depicts the stagnation of the South African economy that started in the second half of the eighties.

Table 8.8 Calculation of the change in the purchasing power of the per capita GDP in the RSA for 1980 to 1990

Year	GDP at current prices (R million)	CPI 1990 = 100	Total population of the RSA (thousands)	Per capita GDP (R)	Per capita GDP in the rand of 1990
1980	56 379	25.5	25 081	2 248	8 816
1981	66 265	29.4	25 618	2 587	8 798
1982	73 877	33.7	26 170	2 823	8 377
1983	84 409	37.8	26 733	3 157	8 352
1984	98 284	42.2	27 308	3 599	8 528
1985	122 448	49.1	27 895	4 390	8 940
1986	129 487	58.2	28 459	4 550	7 818
1987	149 931	67.6	29 025	5 166	7 642
1988	178 077	75.8	29 603	6 016	7 937
1989	207 716	87.4	30 193	6 880	7 872
1990	236 460	100.0	30 797	7 678	7 678

7 Mathematical properties of indices

In the previous sections a variety of unweighted and weighted composite indices were discussed. Due to the complicated nature of some of these indices, their use is limited, while the use of others is restricted by practical considerations. Objective theoretical criteria with which the ideal index should comply have been developed to facilitate the selection of the best index for a certain situation. Two of these are discussed briefly.

The **time reversal test** requires that

$$P_{0,n} \cdot P_{n,0} = 1$$

should apply for an index P, where $P_{0,n}$ and $P_{n,0}$ are given as ratios.

If an index for 1990 with 1980 as the base year is 200, the time reversal test requires the same index for 1980, with 1990 as the base year, to be 50.

The **factor reversal test** requires

$$P_{0,n} \cdot Q_{0,n} = \frac{\sum p_n q_n}{\sum p_0 q_0}$$

where $P_{0,n}$ is a price index and $Q_{0,n}$ is the corresponding quantity index, and both are given as ratios.

A specific index therefore complies with the factor reversal test if the product of the price and quantity index representations is equal to the simple composite value index.

To illustrate, we shall test whether the unweighted composite price index and the Irving Fisher index comply with these two criteria.

EXAMPLE 8.6

SIMPLE COMPOSITE INDEX

Time reversal test:

$$P_{0,n} \cdot P_{n,0} = \frac{\sum p_n}{\sum p_0} \cdot \frac{\sum p_0}{\sum p_n}$$
$$= 1 \quad \text{(does comply)}$$

Factor reversal test:

$$P_{0,n} \cdot Q_{0,n} = \frac{\sum p_n}{\sum p_0} \cdot \frac{\sum q_n}{\sum q_0}$$
$$\neq \frac{\sum p_n q_n}{\sum p_0 q_0} \quad \text{(does not comply)}$$

IRVING FISHER INDEX:

This price index as a ratio is

$$\sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \cdot \frac{\sum p_n q_n}{\sum p_0 q_n}}$$

Time reversal test:

$$P_{0,n} \cdot P_{n,0} = \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \cdot \frac{\sum p_n q_n}{\sum p_0 q_n}} \cdot \sqrt{\frac{\sum p_0 q_n}{\sum p_n q_n} \cdot \frac{\sum p_0 q_0}{\sum p_n q_0}}$$
$$= 1 \text{ (does comply)}$$

Factor reversal test:

$$P_{0,n} \cdot Q_{0,n} = \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \cdot \frac{\sum p_n q_n}{\sum p_0 q_n}} \cdot \sqrt{\frac{\sum q_n p_0}{\sum q_0 p_0} \cdot \frac{\sum q_n p_n}{\sum q_0 p_n}}$$
$$= \frac{\sum p_n q_n}{\sum p_0 q_0} \quad \text{(does comply)}$$

Except for the Irving Fisher index, none of the indices discussed in this chapter comply with both criteria. For this reason the Irving Fisher index is also called the 'ideal index'. A few other criteria are sometimes listed in the literature as requirements for a good index, but the time and factor reversal tests, which are commonly regarded as the most important criteria, will suffice here.

Exercises

1 The value of the exports and imports of the RSA (R million) for the years 1986 to 1990 was as follows (obtained from *SA Statistics 1992*):

	1986	1987	1988	1989	1990
Exports	25 602	25 570	29 809	39 442	42 859
Imports	26 880	28 705	39 428	44 446	44 125

(a) With 1986 as the base year, calculate value indices for exports and imports for the years 1987, 1988, 1989 and 1990. Interpret the results.
(b) Calculate index chains for exports and imports.

2 The economically active population in three industrial divisions in the RSA for three consecutive census years was as follows (obtained from *SA Statistics 1992*):

Sector	Year	Total	Male	Female
Manufacturing	1980	1 465 087	1 105 402	359 685
	1985	1 379 518	986 453	393 065
	1991	1 260 324	880 263	380 062
Financing, insurance,	1980	286 869	159 428	127 441
fixed property and	1985	339 204	173 930	165 274
business services	1991	444 598	227 355	217 243
Community,	1980	2 005 463	871 368	1 134 095
social and personal	1985	1 965 040	718 937	1 246 103
services	1991	2 339 345	852 816	1 486 529

(a) Calculate quantity indices for 1991 and 1985 with 1980 as the base year for all three industrial divisions with regard to both sexes as well as the total.
(b) Discuss the indices with special reference to differences between men and women in respect of the different divisions.

3 The table below indicates the mean retail prices (R/kg) in Pretoria for five types of meat during November for the years 1988, 1989 and 1990. (Obtained from *Statistical News Release P 0141.2* of the Central Statistical Service.)

Type of meat	Nov 1988	Nov 1989	Nov 1990
Beef loin, Super A	14.63	13.70	14.44
Lamb loin, Super	16.20	13.40	13.40
Pork chops, Grade 1	10.23	10.55	10.48
Boerewors	8.95	8.51	9.14
Chicken, cleaned	4.64	4.70	5.48

(a) With November 1988 as the base period, calculate, for November 1989 and November 1990, the simple composite price index P_1.
(b) With November 1989 as the base period, calculate the arithmetic mean of relative prices P_2 for November 1990.

4 Using the prices in Exercise 3 and the weights for beef, mutton, pork, poultry and boerewors given in Appendix C, calculate the following indices for November 1990 (November 1988 = 100):
(a) The weighted composite price index P.
(b) The weighted arithmetic mean of relative prices, P_{rg}.

5 A watch retailer sells mainly four types, namely wrist-watches, alarm clocks, battery-operated wallclocks and luxury wooden grandfather clocks. The mean price (rand) and annual quantities sold in 1985 and 1990 were as follows:

Item	Prices		Quantities sold	
	1985	1990	1985	1990
Wrist-watches	193.12	272.50	926	1 362
Alarm clocks	23.83	38.21	422	511
Battery-operated wall clocks	35.80	73.01	133	381
Luxury wooden grandfather clocks	2 477.20	3 641.81	23	48

Using 1985 as the base year, calculate
(a) $P_L, P_P, P_D, P_F, P_1, P_2, P_3$.
(b) $Q_L, Q_P, Q_D, Q_F, Q_1, Q_2, Q_3$.
(Use the series $\{p_0\}$ and $\{p_n\}$ as the weight series in (b).)

6 The consumer price index for a certain European country was as follows for seven consecutive years:

1985	1986	1987	1988	1989	1990	1991	(1980 = 100)
140.2	154.5	167.8	186.1	202.1	217.2	239.6	

(a) Reduce this series to a series with 1990 = 100.
(b) Compare the series obtained in (a) with the SA Consumer Price Index over the same years as given in Table 8.8. What do you conclude from this?

7 Use Table 8.5 to calculate the inflation rate for the months of January 1989 to December 1990.

8 Piet van Wyk has worked for a large construction firm since 1988. His annual salary in rand (as at April of each year) was as follows:

1988	1989	1990	1991	1992
30 250	40 500	52 600	60 500	75 000

Use the CPI values in Table 8.7 to calculate his deflated salary for each of the years in
(a) 1990 rands; and
(b) 1988 rands.

9 The table below indicates the total number of deaths in road accidents and the number of licensed vehicles (millions) in South Africa during a number of years (obtained from *SA Statistics 1992*):

Year	1984	1985	1986	1987	1988	1989	1990
Deaths	9 621	8 972	9 343	9 905	10 691	10 877	11 159
Number of vehicles	5.388	5.515	5.622	5.719	5.755	5.961	6.145

(a) Calculate quantity indices for 1990 with regard to 1984 for the number of deaths as well as the number of registered vehicles.

(b) Calculate the index chain $\{Q_{84,85}; Q_{85,86}; Q_{86,87}; Q_{87,88}; Q_{88,89}; Q_{89,90}\}$ for the number of deaths and for the number of registered vehicles.

(c) What conclusions can be drawn with regard to the change in the number of deaths and the change in the number of registered vehicles in South Africa during the years 1984 to 1990?

10 With reference to Example 8.3, we consider the April 1984 prices (R1 000) and the number of new cars registered (thousands) during the first four months of 1984 with respect to the seven car makes as given in the table below:

Make	Price April 1984	Number registered Jan–April 1984
1. Audi	20.63	1.330
2. BMW	23.45	5.020
3. Ford	11.88	12.113
4. Mercedes Benz	25.29	4.740
5. Opel	9.71	6.616
6. Toyota	10.29	17.550
7. Volkswagen	10.04	7.590

(a) With April 1984 as the base period, calculate for April 1991,
P_L, P_P, P_D, P_F, P_E, P_1, P_2, P_3 en P_4.

(b) With 1984 as the base period, calculate the following quantity indices for 1991:
Q_L, Q_P, Q_D, Q_F, Q_E, Q_1, Q_2, Q_3 en Q_4.

11 The share prices (rand) and total number of negotiable shares (thousands) of four of the most important companies in a specific sector were as follows:

Company	Price per share		Number of shares	
	June 1990	June 1992	June 1990	June 1992
A	1 320	1 530	20.0	35.0
B	520	330	5.4	2.4
C	4 682	6 310	33.2	41.3
D	2 131	2 020	8.6	11.3

With June 1990 as the base month, calculate for June 1992:
P_L, P_P, P_D, P_F en P_3.

12 With reference to Exercise 6, assume that the consumer price index of the same country with 1991 as the base year for the years 1991 to 1993 was as follows:

1991	1992	1993	(1991 = 100)
100	112.3	127.2	

(a) Obtain one index series for the year 1985 to 1993 with 1990 as the base year.
(b) Obtain the same index series as in (a) with 1991 as the base year.

13 Study the table below with the annual salaries (penultimate notch on salary scale) of two public service posts as at 1 April for the years 1988, 1989 and 1990.

Year	First horticulturist	Museum assistant
1988	R25 371	R12 819
1989	R29 205	R14 751
1990	R32 127	R19 713

Calculate the deflated salaries in terms of
(a) 1990 rands; and
(b) 1988 rands.
Discuss the salary changes in the two posts over the three years and in relation to each other.

14 Use Table 8.5 to calculate the monthly 'inflation rate' with respect to food only for 1991 to 1993. Compare these values with the inflation rate for South Africa (Table 8.6) in the corresponding years.

15 In contrast to the BER index that measures the change in general building costs, the Building Industry Advisory Council (BIAC) index is a measure of change in building costs after the contract date of a building project. The table below indicates this index for 1990 and 1991 (1975 = 100):

Year	January	February	March	April	May	June
1990	777.7	790.4	792.6	800.3	820.1	820.6
1991	890.2	901.0	901.7	903.7	917.9	919.0
Year	**July**	**August**	**Sept**	**Oct**	**Nov**	**Dec**
1990	827.4	839.0	841.7	853.8	874.6	876.0
1991	922.0	933.2	935.2	874.9	885.7	891.2

Do an analysis to compare the increase in building costs during 1990 and 1991 with the increase in the prices of merchandise and services in the corresponding period.

16 The world population (millions), in total and according to continents, was as follows according to the *Statistical Yearbook* of UNESCO (1992):

	1970	1975	1980	1985	1990
World total	3 674	4 052	4 428	4 824	5 248
Africa	366	414	480	557	641
America	509	556	613	665	722
Asia	2 078	2 333	2 562	2 809	3 069
Europe	702	728	750	769	790
Oceania	19	21	23	24	26

(a) Using 1970 as the base year, calculate indices to describe the population growth of the world and of each continent in the years 1975, 1980, 1985 and 1990. Discuss the growth for the various continents.

(b) Estimate the world population as well as the population of Africa for the year 2010. (Note: different methods are possible. Use the analysis procedures explained in Chapters 4, 6 and 7.)

17 The table below is an extract from Chapter 25 of *SA Statistics, 1992* for three main centres of South Africa. Use index figures to discuss the most important differences between the centres with respect to all the meteorological variables in the table by referring to the annual averages but also to seasonal variations.

Meteorological statistics

Main centres – Monthly averages

Month	Temperature °C				Rainfall		
	Mean of		Extreme		Daily average no. hours sunshine	Average total rainfall in mm	Average number of days with rain
	Daily maximum	Daily minimum	Maximum	Minimum			
				East-London			
J	25.2	17.9	32.8	12.6	7.4	75.7	13
F	25.6	18.3	37.4	11.0	7.3	77.5	11
M	24.7	17.5	34.6	11.7	6.8	99.2	12
A	23.5	15.1	34.6	7.2	7.1	68.8	9
M	22.6	12.8	34.9	4.4	7.2	47.6	7
J	20.9	10.5	31.9	4.1	7.4	35.1	5
J	21.0	10.2	33.1	2.8	7.7	28.9	4
A	21.3	11.2	36.7	3.1	7.4	41.5	6
S	21.4	12.5	41.3	5.9	7.2	97.4	9
O	21.6	14.2	39.6	8.3	6.8	111.0	13
N	22.8	15.4	36.6	9.6	7.5	94.3	11
D	24.0	16.7	34.4	11.1	7.7	87.5	13

Month	Temperature °C				Rainfall		
	Mean of		Extreme				
	Daily maximum	Daily minimum	Maximum	Minimum	Daily average no. hours sunshine	Average total rainfall in mm	Average number of days with rain
Johannesburg							
J	26.3	14.3	33.6	5.7	8.4	149.6	15
F	25.6	13.9	33.4	6.7	8.3	129.4	12
M	24.3	12.8	31.1	4.4	8.0	110.4	12
A	22.1	10.3	28.6	–1.1	8.1	47.7	7
M	19.1	6.9	26.1	–4.0	8.8	24.5	4
J	16.5	4.1	24.4	–5.4	8.8	6.5	1
J	16.4	4.1	23.3	–5.6	9.2	10.5	2
A	19.8	6.5	25.9	–5.0	9.8	10.1	2
S	22.8	9.1	32.4	–3.2	9.6	25.6	4
O	25.0	11.8	32.3	0.0	9.0	65.0	9
N	25.3	12.6	33.4	2.1	8.3	126.5	13
D	26.1	13.7	35.0	6.6	8.4	141.1	15
Cape Town (Wingfield)							
J	25.6	15.5	37.4	6.7	11.1	13.5	4
F	26.2	15.5	37.9	5.1	10.4	10.1	3
M	25.0	14.2	39.3	5.4	9.0	16.3	5
A	22.4	11.8	38.8	3.5	7.3	53.9	9
M	19.7	9.5	35.1	–0.7	6.0	78.2	11
J	18.3	7.7	29.7	–1.8	5.9	82.8	11
J	17.0	7.2	28.7	–2.1	6.0	91.8	14
A	17.8	7.9	31.7	–0.6	6.7	69.5	13
S	18.5	9.2	33.9	–0.4	7.5	44.0	9
O	21.0	11.2	32.1	1.3	8.9	29.5	8
N	23.0	13.0	33.9	4.4	10.0	20.8	6
D	24.5	14.3	37.6	5.3	11.0	10.2	4

18 The table below appeared in a Pretoria English afternoon paper on 27 August 1992 and reflects the prices of a representative basket of groceries at five well-known supermarkets in Pretoria as observed on a certain day:

Shop-a-round

This is another in a series of monthly price checks of popular items in five of the city's leading supermarkets. All the shops used for the survey belong to well-known supermarket chains. Prices were noted on Tuesday and are subject to fluctuations, so they do not necessarily apply today.

	OK Glenfair	Hyperama Menlyn	Checkers Lynnwood	Checkers Waterkloof	Pick 'n Pay Lynnwood
Meat:					
T-bone steak, R/kg	15.99	18.48	15.78	17.98	17.49
Tenderised steak, R/kg	17.99	18.68	16.78	12.97	16.98
Topside roast, R/kg	16.95	17.95	15.78	13.49	16.98
Pot roast, R/kg	9.99	9.49	9.98	11.99	10.58
Lamb braai chops, R/kg	9.99	10.99	11.97	11.97	17.86
Lamb neck, R/kg	11.99	14.99	12.90	12.98	15.86
Leg of lamb, R/kg	14.99	14.99	14.98	13.98	15.86
Stewing beef, R/kg	7.99	8.98	7.77	7.77	9.98
Dairy:					
Rama, 500g	2.79	2.99	3.39	3.39	3.29
Cheddar cheese, R/kg	17.85	16.79	18.39	16.59	16.99
Bliss yoghurt, 500 ml	3.09	2.89	2.68	2.68	2.19
Vegetables and fruit:					
Bananas, R/kg	1.69	1.89	2.19	2.19	2.09
Potatoes, R/kg	2.59	3.35	1.99	2.09	2.79
Gems, R/kg	2.29	3.59	3.99	3.59	3.29
Onions, R/kg	3.69	3.74	2.39	2.59	3.29
Tomatoes, R/kg	2.99	2.99	3.29	2.99	3.99
Dry foods:					
Sugar, 2,5 kg	4.89	5.29	5.39	5.19	5.49
Weetbix, 500g	3.65	2.99	3.69	3.69	3.19
Cake flour, 2,5 kg	9.99	9.99	10.39	10.39	9.99
Rice, 2 kg	4.99	4.49	5.60	3.88	4.79
Knorr soup (1 pk)	0.79	0.89	0.99	0.99	0.89
Marmite, 250 g	7.25	–	6.49	6.49	6.29
Household cleaners:					
Omo, 1 kg	7.69	6.99	6.79	6.79	6.59
Country Pride fabric softener, 2 ℓ	3.99	3.99	5.39	5.39	4.99
Handy Andy, 750 ml	3.19	3.19	3.19	3.19	3.19
Other:					
Ricoffy, 750 g	8.99	9.99	8.89	8.89	9.99
Trinco tea bags, (100 tagless)	4.89	5.69	5.69	5.69	5.59
Lyles syrup, 1 kg	4.99	5.59	4.69	5.99	4.69
Baked beans in tomato sauce, 410 g	1.29	1.29	1.15	1.49	1.39
Dogmor dogfood, 10 kg	17.99	19.99	18.29	18.29	19.99

(a) Use the weights of commodities in Appendix C and arrange the five supermarkets in sequence from the cheapest to the most expensive.

(b) What is the percentage price difference between the cheapest and the most expensive supermarket?

CHAPTER 9

Introduction to probability theory

1 Introduction

The word 'probability' is used colloquially in many contexts. An economist, for example, can predict that, 'in all probability', an economic boom will continue for a particular period of time, or sport enthusiasts can discuss the 'chances' of their team winning a match, while the weather bureau can forecast the 'possibility' of rain. Although the terms 'probability', 'chance' and 'possibility' are common usage, many people do not understand these terms, other than when they are used in a relative sense. Everyone knows, for instance, what we mean if we say that one project has a better chance of success than another. However, when we have to quantify the respective chances of the two projects succeeding, assessments are based on intuition.

As a concept, probability has a vague and even mystic meaning for uninformed people, but it is actually based on a sound mathematical theory and can be a powerful aid in making important decisions. In modern society, where daily decisions in virtually every field necessitate risks that involve large sums of money or human lives, it is essential to understand and use the theory of probability correctly. Thus, for example, financial institutions are constantly having to make decisions about expenditure on building projects, shares and other investments, while doctors frequently have to make snap decisions regarding choice of specific treatments for patients. Such decisions by their very nature involve a degree of uncertainty and risk. The only scientific support for such decisions is from statistical techniques that are based on accurate information already collected. In this regard probability theory plays an extremely important role.

In the first part of this book we concentrated on **descriptive statistics**, while in the second part, the focus is on **statistical inference**, that is, statistical generalization. These two aspects of statistics are not separate, in fact it is impossible to have statistical inference without descriptive statistics, and descriptive techniques are usually applied as a first step in the process of generalization. Probability theory is an independent field of study for mathematical statisticians and ongoing research is being carried out in this field. However, since the purpose of this book is to place the application of statistics within reach of students and researchers in all disciplines in a simple

manner, we shall not delve too deeply into the theoretical basis of probability theory.

In this chapter the concept 'probability' is defined and the calculation of probability is illustrated, while certain rules of probability will also be discussed. In Chapter 10 we look at the concept 'theoretical distribution', and certain special distributions and their characteristics are discussed. With this background knowledge the reader should be able to understand statistical inference, which is dealt with in the subsequent chapters.

2 Basic concepts

When an insurance company sells a car insurance policy to a client, the company is unable to predict whether the client will claim against the policy within the next twelve months. In order to determine a premium that will be acceptable to the client and also provide the company with adequate financial remuneration for the risk it has to carry on behalf of the client, the company makes an estimate of the probability that the client will claim and of the expected size of such a claim. Suppose, for example, the insurance company knows from previous experience that from 20 000 similar clients, normally about 1 800 claim and that the average size of the claim is R8 550. They then set the probability that the specific client will claim at $\frac{1\,800}{20\,000} = 0.09$ or 9%, and if the client does claim, the expected claim size is R8 550. Strictly speaking, the fraction 0.09 is not the probability that the client will submit a claim, but an approximation of the **relative frequency** of similar policy holders who will submit a claim against the policy concerned over a period of twelve months. This approximate relative frequency normally gives a sufficient indication of the required probability.

In the above example it is clear that the probability of a certain event can be approximated by the relative number of times the event has occurred in a large number of similar situations. The so-called **frequency limit definition** of probability flows from this approach. Let us suppose a certain event is indicated by the symbol A and the probability that A will occur (be realized), by $P(A)$.

Suppose an event A is realised x times out of a possible n times and x/n indicates the relative frequency of the number of realisations of A. According to the **frequency limit definition**, the probability of A is then given by

$$P(A) = \lim_{n \to \infty} \frac{x}{n}$$

Using the above example, we can then say

$$P(\text{policy holder submits claim}) = \lim_{n \to \infty} \frac{\text{number of policy holders submitting claims } (x)}{\text{total number of policy holders } (n)}$$

$$\approx \frac{1\,800}{20\,000}$$
$$= 0.09$$

In the above example accurate information already collected enabled us to obtain at least a good approximation of the probability that a client will claim in the policy term. However it should be emphasized that the exact value of the required probability can be determined in terms of the frequency limit definition only if the number of similar clients is 'infinite'.

The study of probability originated with individuals who made money from gambling and realized that a mathematical analysis of such games would give them an advantage over their competitors. Mathematicians were called in and they not only solved the required problems, but also started working on more complicated ones. Today textbooks explain probability on the basis of games of chance, not because probability is rooted in gambling, but because many other applications of probability can best be explained on the basis of simple games of chance.

The simplest game of chance is tossing a coin. Various scientists have repeatedly tossed a coin and, using the above definition, have tried to calculate the probability of heads accurately. During the Second World War while the English mathematician, J.E. Kerrich, was interned, he tossed a coin 10 000 times and the result was:

$$P(\text{heads}) \approx \frac{5\,067}{10\,000}$$
$$= 0.5067$$

In modern times such an experiment would be regarded as a waste of time. However, consider the following experiment that was conducted recently.

EXAMPLE 9.1

An experiment is conducted to determine the probability of heads coming up with a buckled 1c coin. The coin is tossed fifty times and the number of heads is noted.

This experiment is carried out ten times and after each set of fifty attempts the relative frequency of the number of heads is recorded for the fifty repetitions as well as for all attempts completed at that point. The results are summarized in Table 9.1 and the relative frequency of all completed attempts is represented graphically in Figure 9.1. (The only reason for drawing the straight lines between

the points plotted in the figure was to facilitate the study of the figure.) This table and figure therefore reveal that

P (heads with buckled coin) ≈ 0.51.

Table 9.1 Relative frequencies of the number of heads in ten experiments in which a 1c coin was tossed 50 times per experiment. The relative frequency for all completed attempts is also indicated

Number of experiment	Relative frequency of 50 attempts	Relative frequency of all completed attempts
1	$\frac{31}{50} = 0.62$	$\frac{31}{50} = 0.62$
2	$\frac{22}{50} = 0.44$	$\frac{53}{100} = 0.53$
3	$\frac{29}{50} = 0.58$	$\frac{82}{150} = 0.55$
4	$\frac{28}{50} = 0.56$	$\frac{110}{200} = 0.55$
5	$\frac{22}{50} = 0.44$	$\frac{132}{250} = 0.53$
6	$\frac{27}{50} = 0.54$	$\frac{159}{300} = 0.53$
7	$\frac{23}{50} = 0.46$	$\frac{182}{350} = 0.52$
8	$\frac{23}{50} = 0.46$	$\frac{205}{400} = 0.51$
9	$\frac{23}{50} = 0.46$	$\frac{228}{450} = 0.51$
10	$\frac{26}{50} = 0.52$	$\frac{254}{500} = 0.51$

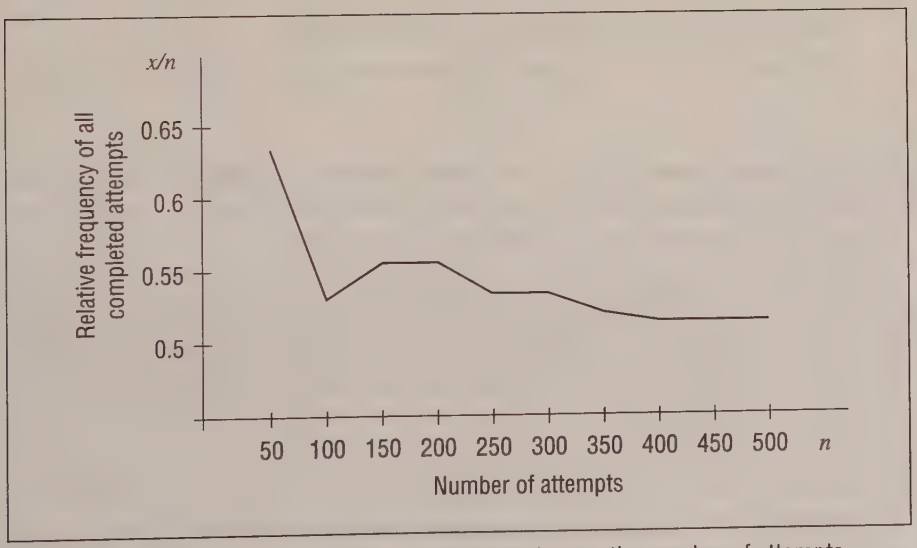


Figure 9.1 Relative frequency of the number of heads over the number of attempts

From the above example and from other similar experiments it is clear that $\frac{x}{n}$, the relative frequency of all completed attempts, fluctuates considerably, but later stabilizes around a fixed value. When a coin is tossed the result is not known until it is done. This is characteristic of games or experiments of chance. Such an experiment is known as a **stochastic experiment**. In all such **stochastic phenomena** there is a **statistical regularity**. If the same coin is tossed again 500 times, the relative frequency $\frac{x}{n}$ will not differ much from 0.51.

The frequency limit definition is important as it eliminates some of the vagueness of the probability concept. It also enables us to determine approximate probabilities for certain stochastic phenomena in an experimental (empirical) manner, as illustrated in Example 9.1.

The approximate determination of probability by means of the frequency limit definition is important in, for example, insurance, but this is only one of the ways in which probabilities can be allocated to certain events. In many cases probabilities are allocated on the basis of **subjective** considerations and not on any stochastic experiments. An example can be found where the board of directors of a company considers the 'chances' of project A succeeding if a large amount of capital is invested in it. Probably the best-known example of subjective allocation of probabilities in South African society is the daily or even long-term weather forecast. If there is a '60% chance' of general rain over the interior, this forecast cannot be tested by means of the frequency limit definition, and the public has to assume that the Weather Bureau has quantified the probability of rain on the basis of other considerations (that are scientifically accountable).

In many situations the probability of a certain event can be determined from the intrinsic characteristics of the experiment concerned. Examples are the following:

(i) The probability of a six with a balanced dice is $\frac{1}{6}$.

(ii) The probability of a hearts card from a full deck of 52 cards is $\frac{1}{4}$ (provided the cards were well shuffled beforehand).

(iii) The probability of a red card with the number 10 on it from a deck of 52 cards is $\frac{2}{52}$.

It is thus clear that, in addition to the above methods of **relative frequency** and **subjective allocation**, there is also a third method of allocating probability to events. This is summarized in the following definition.

If a stochastic experiment has n different equally possible outcomes and m of these are favourable for realizing the event, A, the probability of A in terms of the **classical definition of probability** is given by

$$P(A) = \frac{m}{n}$$

It must be emphasized, however, that the classical definition can be applied only if all the results are equally possible. Thus, for example, the definition

cannot be used to calculate the probability of heads with the buckled coin in Example 9.1. The application of the frequency limit definition was the only possible solution there.

EXAMPLE 9.2

An auditing firm has limited time to check the cash payment transactions of a large company before reporting to the board of directors. From previous experience with similar companies the auditors know that all such transactions can be divided into three categories, namely:

T: cheques
K: credit cards
D: debit orders

It is also known that the relative frequencies of the three types of transactions with this company are approximately the following:

T: Relative frequency = 0.62
K: Relative frequency = 0.27
D: Relative frequency = 0.11

Due to the time restriction the auditors decide to draw random samples of transactions from the total pool of payment transactions.
The actual situation can now be copied or simulated by a simple game of chance. A container with 100 cards that are identical in all regards except colour is used as a model. Sixty-two of the cards are green, 27 are red and 11 are yellow. The green cards represent cheque payments (T), the red cards credit card payments (K) and the yellow cards debit orders (D). If the cards are shuffled well and one card is drawn at random without looking, it follows immediately from the classical definition of probability that

$$P(T) = P(\text{green card drawn}) = \frac{62}{100} = 0.62$$

$$P(K) = P(\text{red card drawn}) = \frac{27}{100} = 0.27$$

$$P(D) = P(\text{yellow card drawn}) = \frac{11}{100} = 0.11$$

Suppose we then repeatedly draw one card from the container, note the colour and replace it, we can expect the following from the frequency limit definition after many such repetitions:

$P(T) \approx 0.62$
$P(K) \approx 0.27$
$P(D) \approx 0.11$

Suppose, in the same example, we need to calculate the probability of obtaining no more than 10 cheque payments from a random sample of 50 payment transactions, or no debit order from a random sample of 20 transactions, et cetera,

then the solution to the problem is more complicated. However, we can determine these probabilities by using the rules of probability. These rules emanate from the third definition of probability, which is known as the **axiomatic definition**. **Set theory** is used to make this definition understandable.

In terms of the axiomatic approach, any event occurring during an experiment is represented as a set of descriptions of favourable outcomes for the event. Take, for example, the following stochastic experiment. A random digit is drawn from a table of random numbers (for example Table B.1). The digits 0, 1, 2, ..., 9 are arranged randomly in such a way that each one appears in the sequence approximately an equal number of times without any pattern appearing in the sequence. If these digits are used as discussed in Chapter 2, Section 4, it follows that any one digit has an equal chance of being drawn, i.e.

$$P(\text{digit } 0) = P(\text{digit } 1) = \ldots = P(\text{digit } 9) = \frac{1}{10}$$

Now take the following event:

A: odd digit drawn

A can then be represented as follows in set notation:

$$A = \{1, 3, 5, 7, 9\}.$$

Note that the experiment consists of drawing one random digit and the **outcome** is described by a specific symbol, for instance '7' or '8'. If this symbol or description is an element of A, we say that 'A has realized'. In other words, if we draw one random digit and find a '7', A has taken place, but if an outcome with the description '8' is found then A has not taken place. The universal set with descriptions of all possible outcomes is known as the **sample space** S. In the above examples the sample space is

$$S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}.$$

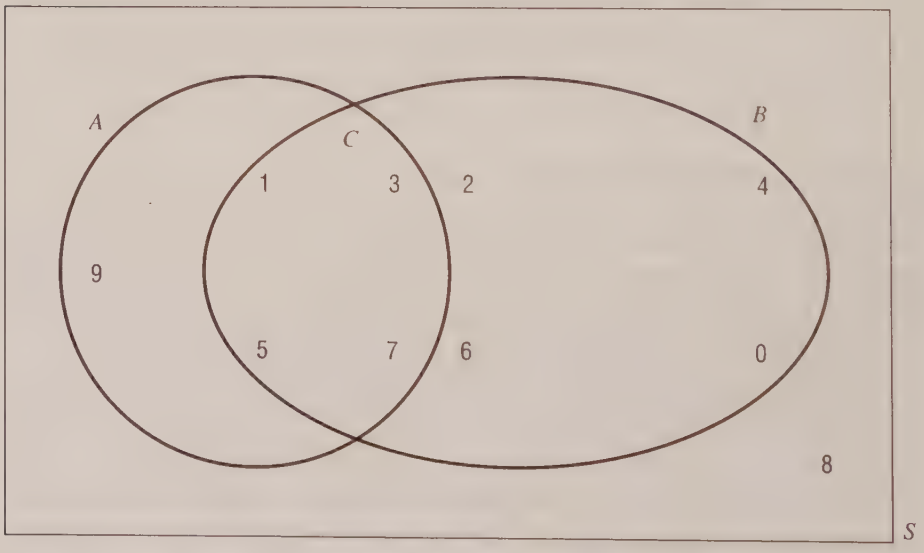


Figure 9.2 Venn diagram of three events, A, B and C

The sample space is also known as the **certain event** because we are certain, when conducting this specific experiment, that the event will take place. Since events can be regarded as subsets of the certain event S, the notation and operations used in set theory are directly applicable to probability. Events can, for example, be represented in a Venn diagram. In Figure 9.2 a Venn diagram is drawn for the following events:

A: odd digit drawn
B: digit smaller than 8 drawn
C: odd digit smaller than 8 drawn.

Table 9.2 gives the explanation and Venn diagram representations of the events that are of specific importance to probability.

Table 9.2 Meaning and Venn diagram representations of events

Event	Meaning	Venn diagram
S	The certain event takes place	
$\varnothing$	The impossible event	
$\bar{A}$	The event A does not take place, i.e. A complement takes place	
$A \cup B$	A and/or B takes place, i.e. at least one of the two events is realized	
$A \cap B$	A and B take place, i.e. A and B are realized	
$A \cap B = \varnothing$	A and B are mutually exclusive events, i.e. A and B cannot take place simultaneously	

As already stated, the rules of probability can be derived from the axiomatic definition and for this reason this definition is now given in terms of set theory.

According to the **axiomatic definition** of probability, the probability of any event A is non-negative, the probability of the certain event S is equal to 1 and the probability of the realisation of A or B is equal to the probability of A plus the probability of B, provided A and B are mutually exclusive events.

It is important to point out that

the frequency limit definition,
the classical definition and
the axiomatic definition

of probability neither contradict one another nor represent different views of the concept of probability. In fact, it can be shown that the classical definition of probability meets the requirements of the axiomatic definition and that both definitions can be confirmed empirically by applying the frequency limit definition.

3 Rules of probability

As the formal derivation of the rules of probability is beyond the scope of this book, a discussion of these rules on the basis of examples will suffice. All the rules can be justified intuitively and are easily understood with the use of Venn diagrams.

3.1 The boundaries of probability

If A is the impossible event $\varnothing$, then $P(A) = 0$. On the other hand, if A is the certain event S, $P(A) = 1$. It is clear that the probability of any event has to lie between these two boundary values.

For any event A, it is therefore true that

$$0 \leq P(A) \leq 1.$$

Note that if probability is regarded as the limit of a relative frequency or as the ratio of favourable outcomes to total number of equally possible outcomes, this rule of probability boundaries is self-evident.

3.2 The addition rule

The probability of any two events A and/or B occurring in the same stochastic experiment, namely $P(A \cup B)$, cannot be calculated by simply applying the axiomatic definition.

According to the axiomatic definition,

$$P(A \cup B) = P(A) + P(B)$$

provided A and B are mutually exclusive events. However, consider events A and B when a random digit is drawn from Table B.1:

A: odd random digit with $P(A) = 0.5$
B: random digit smaller than 8 drawn with $P(B) = 0.8$.

Now $P(A) + P(B) = 0.5 + 0.8$
$= 1.3$

which is clearly contrary to rule 3.1 above, if this were to be taken as the value of $P(A \cup B)$.

From Figure 9.3 it is clear that the number 1.3 includes the probability of the overlapping section (the event $A \cap B$) twice. For this reason the correct value of $P(A \cup B)$ will be obtained by subtracting $P(A \cap B)$ from the number again, that is

$$P(A \cup B) = 0.5 + 0.8 - 0.4$$
$$= 0.9$$

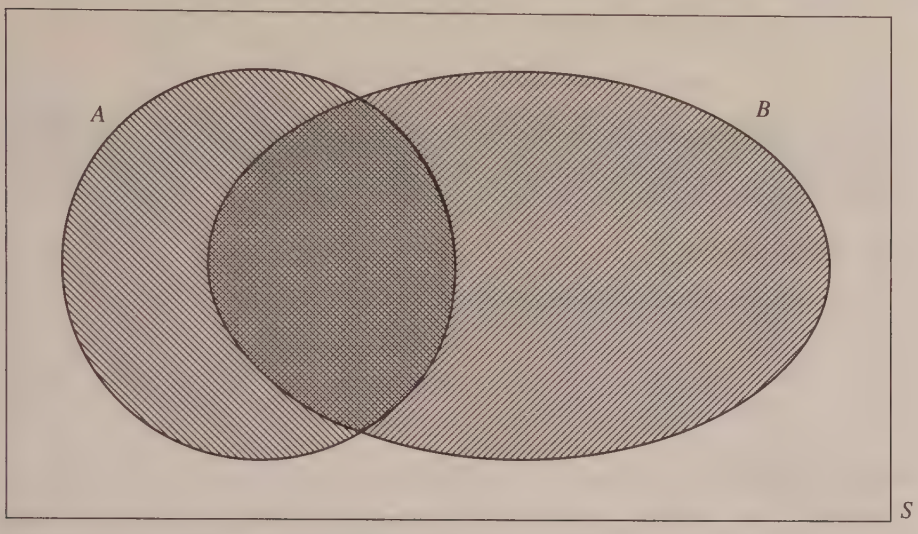


Figure 9.3 Venn diagram for calculating $P(A \cup B)$

The **addition rule** for any two events A and B is expressed as

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

The addition rule can be extended to more than two events by applying the rule for two events. For example, take three events A, B and C. It follows that

$$A \cup B \cup C = A \cup (B \cup C) = A \cup D \text{ with } D = B \cup C.$$

By applying the addition rule to events A and D, it follows after simplification that

$$\begin{aligned} P(A \cup B \cup C) = {} & P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) \\ & - P(B \cap C) + P(A \cap B \cap C). \end{aligned}$$

The addition rule can be extended to any number of events. However, the rule then becomes much more complicated, except in cases where all the events are mutually exclusive. If, for example, $E_1, E_2, \ldots, E_n$ are n **mutually exclusive events**, the addition rule for these events is

$$\begin{aligned} P(E_1 \cup E_2 \cup \ldots \cup E_n) &= P(E_1) + P(E_2) + \ldots + P(E_n) \\ &= \Sigma P(E_i) \end{aligned}$$

3.3 The complement rule

If A does not take place during the execution of a stochastic experiment, $\bar{A}$ does. In other words:

$$S = A \cup \bar{A}$$

Now, in terms of the axiomatic definition,

$P(S) = P(A) + P(\bar{A})$, since $P(A \cap \bar{A}) = 0$

Since $P(S) = 1$, we have derived the following rule:

> The **complement rule** for an event A can be expressed as
> $P(\bar{A}) = 1 - P(A)$.

Thus, for Example 9.2

$$\begin{aligned} P(\text{transaction not a cheque payment}) &= P(\bar{T}) \\ &= 1 - P(T) \\ &= 1 - 0.62 \\ &= 0.38 \end{aligned}$$

EXAMPLE 9.3

According to the records of a motor vehicle insurance company, 1 800 of its clients submitted claims for at least one event during the past year.

Of the clients who submitted claims for one event, 730 claimed for personal damage, 539 for theft, 621 for damage suffered by a third party, and 432 for both personal damage and damage suffered by a third party. The rest of the policy holders submitted claims for more than one event.

If this information is typical of the claim pattern in any year, determine the probability that a randomly selected policy holder who suffers damage during the next year

(a) will claim for damage in one event;
(b) will not claim for theft; and
(c) will claim for own damage or for the damage suffered by a third party in one event.

Also determine

(d) the probability that any policy holder will claim for more than one event.

First define the following events:

A: claim due to own damage in one event
B: claim due to damage suffered by a third party in one event
C: claim due to theft
D: claim due to more than one event

The problem can best be understood if the data is represented in a Venn diagram. In this regard study Figure 9.4, which depicts all the relevant data. Set A comprises $298 + 432 = 730$ elements, set B $189 + 432 = 621$ elements, set C 539 elements and set D $1\,800 - (730 + 621 + 539) = 342$ elements.

(a) $$P(A) = \frac{730}{1\,800}$$
$$= 0.41$$

(b) $$P(\bar{C}) = 1 - P(C)$$
$$= 1 - \frac{539}{1\,800}$$
$$= 0.70$$

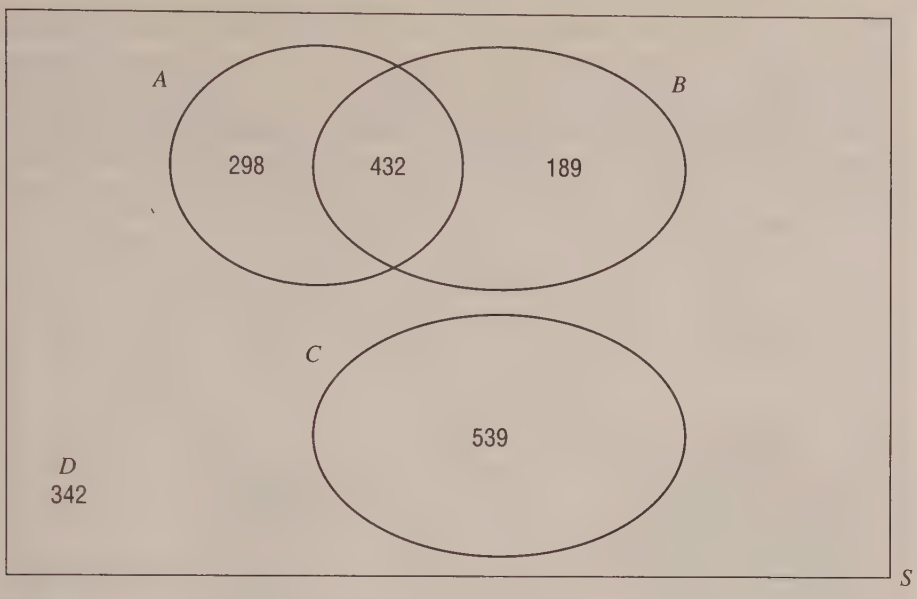


Figure 9.4 Venn diagram of claim pattern for one year

(c) $P(A \cup B) = \frac{730}{1\,800} + \frac{621}{1\,800} - \frac{432}{1\,800}$

$= 0.51$

(d) $P(D) = \frac{342}{1\,800}$

$= 0.19$

3.4 Conditional probability

As already stated, the probability of drawing any one of the random numbers 0, 1, ..., 9 is equal to $\frac{1}{10}$. Now suppose that person X draws a random number and asks person Y: "What is the probability that I have drawn a 7?" Y will naturally answer "one-tenth". However, if X gives away part of the secret, namely that he has drawn an odd number, Y can use the classical definition to calculate that the probability of it being 7 is now $\frac{1}{5}$. Person Y has intuitively applied the rule of conditional probability.

Suppose A and B are two events that can take place during the execution of a stochastic experiment. The **conditional probability** of A, given that B has taken place, is then defined by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \text{ provided } P(B) > 0$$

If $P(A)$ is interpreted in a Venn diagram as the area associated with the event A relative to the area associated with the certain event, the difference between probabilities $P(A)$ and $P(A \mid B)$ can be explained by means of Figure 9.5. In Figure 9.5(a) $P(A)$ can be regarded as the area of the shaded part in relation to the area of S. On the other hand, $P(A \mid B)$ in Figure 9.5(b) can be regarded as the area of $A \cap B$ relative to the area of B, since B plays the role of the 'certain event' in the conditional case.

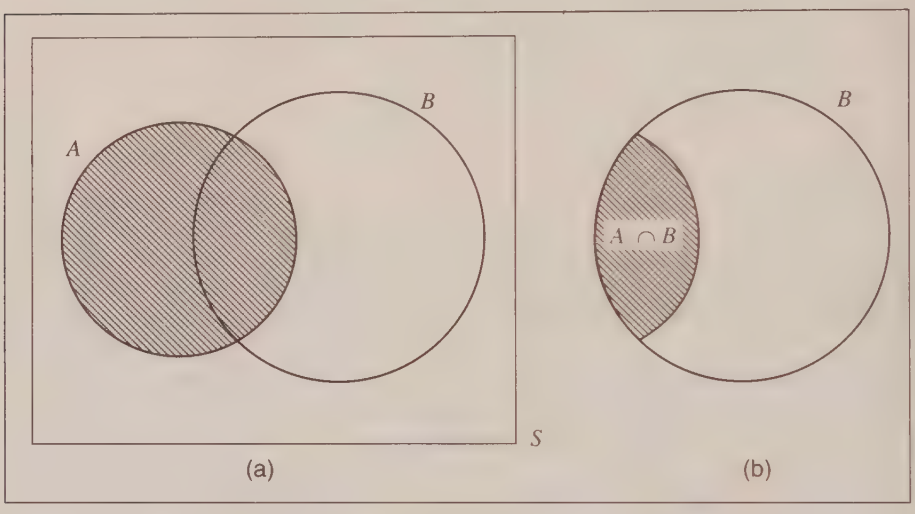


Figure 9.5 Venn diagram of $P(A)$ and $P(A \mid B)$

Note that the (unconditional) probability of A can also be expressed as a conditional probability:

$$P(A) = \frac{P(A \cap S)}{P(S)}$$

$$= P(A \mid S)$$

If the definition of conditional probability is applied to the above dialogue between persons X and Y, then for

A: 7 drawn (with $P(A) = \frac{1}{10}$)

B: odd number drawn (with $P(B) = \frac{5}{10}$)

$A \cap B$: 7 and odd number drawn (with $P(A \cap B) = \frac{1}{10}$)

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{\frac{1}{10}}{\frac{5}{10}}$$

$$= \frac{1}{5}$$

EXAMPLE 9.4

In Example 9.3 A, B and C represent claims that arise from one event (either accident or theft), while D represents claims that arise from more than one event (for example: two accidents or one accident and one theft). The following question can now be asked:

"What is the probability that a client will claim in one event for a third party, given that he claims for own damage in the same event?"

In terms of the notation in Example 9.3, we should therefore determine

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

$$= \frac{\frac{432}{1\,800}}{\frac{730}{1\,800}}$$

$$= 0.59$$

Similarly, it follows that

P(claim due to theft, given claim for exactly one event)

$$= P(C|A \cup B \cup C)$$

$$= P(C|\overline{D})$$

$$= \frac{P(C \cap \overline{D})}{P(\overline{D})}$$

$$= \frac{\frac{539}{1\,800}}{1 - \frac{342}{1\,800}}$$

$$= 0.37$$

3.5 The product rule

Solving $P(A \cap B)$ from the definition of conditional probability leads to the product rule of probabilities.

Suppose A and B are two events that can take place in the execution of a stochastic experiment. The **product rule of probabilities** is given by

$P(A \cap B) = P(B) \times P(A|B)$, provided $P(B) > 0$.

EXAMPLE 9.5

A car dealer has determined by means of market research that 60% of his potential clients prefer the 1800 cc model of a medium-sized car, while 40% prefer the stronger 2000 cc model. From a postal survey questionnaire among his existing clients he obtained the following additional information:

1800 cc model: 35% prefer air conditioning
2000 cc model: 75% prefer air conditioning

If he needs to use the above information when ordering vehicles to exhibit in his showroom, what percentages of the four possibilities should he order?

Let A: Client prefers 1800 cc model
B: Client prefers air conditioning
In terms of probabilities the following now applies:

$P(A) = 0.60$
$P(\bar{A}) = 0.40$
$P(B \mid A) = 0.35$ and $P(\bar{B} \mid A) = 0.65$
$P(B \mid \bar{A}) = 0.75$ and $P(\bar{B} \mid \bar{A}) = 0.25$

Therefore

$$\begin{aligned} P(\text{client prefers 1800 cc model with air conditioning}) &= P(A \cap B) \\ &= P(A)\,P(B \mid A) \\ &= 0.60 \times 0.35 \\ &= 0.21 \end{aligned}$$

$$\begin{aligned} P(\text{client prefers 1800 cc model without air conditioning}) &= P(A \cap \bar{B}) \\ &= P(A)\,P(\bar{B} \mid A) \\ &= P(A)\{1 - P(B \mid A)\} \\ &= 0.6 \times (1 - 0.35) \\ &= 0.39 \end{aligned}$$

Similarly, it follows that

$$\begin{aligned} P(\text{client prefers 2000 cc model with air conditioning}) &= P(\bar{A} \cap B) \\ &= P(\bar{A})\,P(B \mid \bar{A}) \\ &= 0.40 \times 0.75 \\ &= 0.30 \end{aligned}$$

and

$$\begin{aligned} &P(\text{client prefers 2000 cc model without air conditioning}) \\ &= 1 - (0.21 + 0.39 + 0.30) \\ &= 0.10 \end{aligned}$$

The four probabilities 0.21, 0.39, 0.30 and 0.10 give the respective ratio in which the dealer should order the vehicles.

Just as the Venn diagram often simplifies the solution of a problem, problems can also be represented and solved by a **tree diagram**. Thus, for example, the potential clients and their preferences for car models with or without air conditioning, as discussed in Example 9.5, can be represented by means of the tree diagram in Figure 9.6.

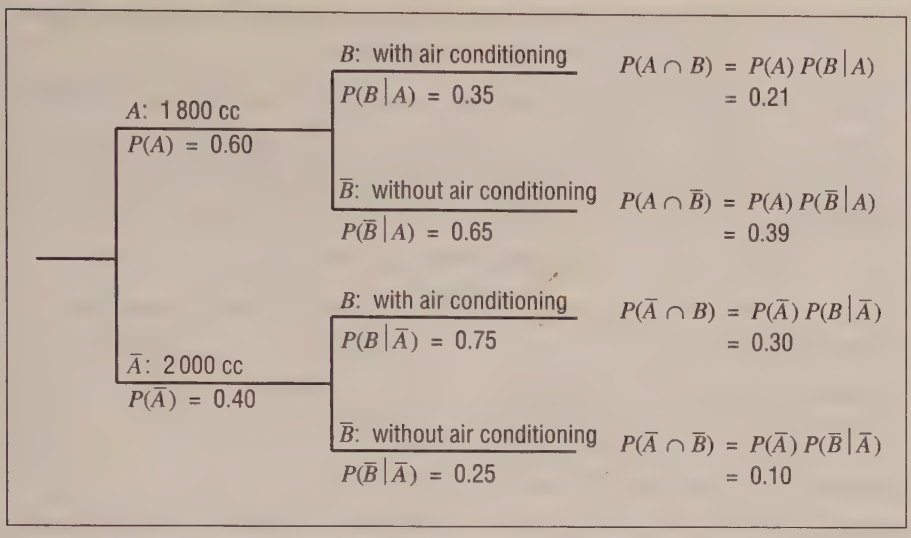


Figure 9.6 Potential clients and their preference for car models with or without air conditioning

Tree diagrams are very useful in probability problems in which the concepts of conditional probabilities and the product rule are used, particularly in more complicated situations. Tree diagrams are also used in counting techniques (see Section 4).

The product rule can also be extended to more than two events. For example, let

$$A \cap B \cap C = (A \cap B) \cap C = E \cap C \text{ with } E = A \cap B$$

By applying the product rule for the two events E and C, it follows after simplification that

$$P(A \cap B \cap C) = P(A) \times P(B|A) \times P(C|A \cap B)$$
provided $P(A) > 0$ and $P(A \cap B) > 0$

The product rule can be extended to any number of events.

3.6 Statistically independent events

In a game of chance with dice in which a player wins if he throws a 6 during his turn, a player who has not won for a long time often expects the chance of his throwing a 6 to increase the longer he tries to throw a 6. The merits of this 'theory' can be assessed if we use the concept of **statistical independence of events**.

Two events A and B are statistically independent if the realization of one does not affect the realization of the other one in any way.
If A and B are two **statistically independent events**, then

$$P(A|B) = P(A) \text{ and } P(B|A) = P(B).$$

From this definition one can deduce that, for statistically independent events A and B the product rule can be simplified as follows:

$P(A \cap B) = P(A) \times P(B)$, provided A and B are statistically independent.

Note that the latter expression does not necessarily imply that A and B are statistically independent (see Exercise 9.19). The product rule for n statistically independent events $F_1, F_2, \ldots, F_n$ can similarly be expressed as follows:

$$P(F_1 \cap F_2 \cap \ldots \cap F_n) = P(F_1)P(F_2) \ldots P(F_n).$$

When a dice is rolled repeatedly, there is no way in which the outcomes can influence one another.

Suppose

A_i: Outcome is a 6 in the ith trial, $i = 1, 2, \ldots, 10$

it follows that these events are statistically independent.

Therefore

$P(\text{outcome is a 6 in the second trial} \mid \text{outcome not a 6 in the first trial})$
$= P(A_2 \mid \bar{A}_1)$
$= P(A_2)$, because A_1, A_2 are independent
$= \frac{1}{6}$

Similarly,

$P(\text{outcome is a 6 in the tenth trial} \mid \text{no outcome of 6 in first nine trials})$
$= P(A_{10} \mid \bar{A}_1 \cap \bar{A}_2 \cap \ldots \cap \bar{A}_9)$
$= P(A_{10})$, because $A_1, A_2, \ldots, A_{10}$ are independent
$= \frac{1}{6}$

The 'theory' that the chances for a 6 'improve' the longer the preceding series of non-sixes lasts, is therefore totally unfounded.

What actually happens here, is that the player confuses two events, namely:

A_n: Outcome is a 6 in the nth trial | outcome not a 6 in the previous $(n-1)$ trials

B_n: At least one 6 in n trials

where the value of n is high. It is clear that

$$P(A_n) = \tfrac{1}{6} \text{ for any } n$$

and
$$P(B_n) = 1 - P(\text{no 6 in } n \text{ trials})$$
$$= 1 - \left(\tfrac{5}{6}\right)^n$$
$$= \begin{cases} 0.838 \text{ for } & n = 10 \\ 0.974 \text{ for } & n = 20 \text{ etc.} \end{cases}$$

EXAMPLE 9.6

The Chief Directorate Civil Aviation of the Department of Transport has to be informed of every aviation incident at every airport in the country. An aviation incident is regarded as any event in which a potentially dangerous situation develops when an aircraft moves under its own power. At a certain airport there are about 84 aviation movements (landings or take-offs) per week and the relative frequency of aviation incidents during take-offs or landings has been established as 0.0001. Determine the probability that no incident will be noted during one week's movements.

Let A_i = incident noted during ith movement.

$P(A_i) = 0.0001$ and $P(\overline{A}_i) = 0.9999$

Since the individual take-offs or landings can be regarded as statistically independent events, the required probability (by applying the product rule for independent events) is

$$\begin{aligned} & P(\overline{A}_1 \cap \overline{A}_2 \cap \ldots \cap \overline{A}_{84}) \\ &= P(\overline{A}_1)P(\overline{A}_2) \ldots P(\overline{A}_{84}) \\ &= (0.9999)^{84} \\ &= 0.99163 \\ &\approx 0.99 \end{aligned}$$

3.7 The law of total probability

Take a population that consists of groups of individuals, within which the relative frequency of the incidence of a certain characteristic differs for the various groups. When an individual is selected at random from this population and it is not known from which group the individual is selected, the probability that the individual does possess the characteristic can be determined with the aid of the **law of total probability**.

This rule of probability involves a combination of the addition rule and the product rule of probabilities and can be formulated as follows:

> Take events $B_1, B_2, \ldots, B_m$ which are all mutually exclusive and for which $B_1 \cup B_2 \cup \ldots \cup B_m = S$, the certain event. Let D be any other event based on the same sample space S. According to the **law of total probability**,
>
> $$\begin{aligned} P(D) &= \sum P(B_i \cap D) \\ &= \sum P(B_i)\, P(D \mid B_i), \text{ provided } P(B_i) > 0,\ i = 1, 2, \ldots, m \end{aligned}$$

The rule is easier to understand if a Venn diagram is drawn for the case of three mutually exclusive events, B_1, B_2 and B_3, and another event D, as given in Figure 9.7. In this figure

$$D = (B_1 \cap D) \cup (B_2 \cap D) \cup (B_3 \cap D)$$

with $B_1 \cap D$, $B_2 \cap D$ and $B_3 \cap D$ mutually exclusive events.
From the addition rule for three mutually exclusive events it follows that:

$$P(D) = P(B_1 \cap D) + P(B_2 \cap D) + P(B_3 \cap D)$$
$$= P(B_1)P(D|B_1) + P(B_2)P(D|B_2) + P(B_3)P(D|B_3)$$

by application of the product rule of probability.

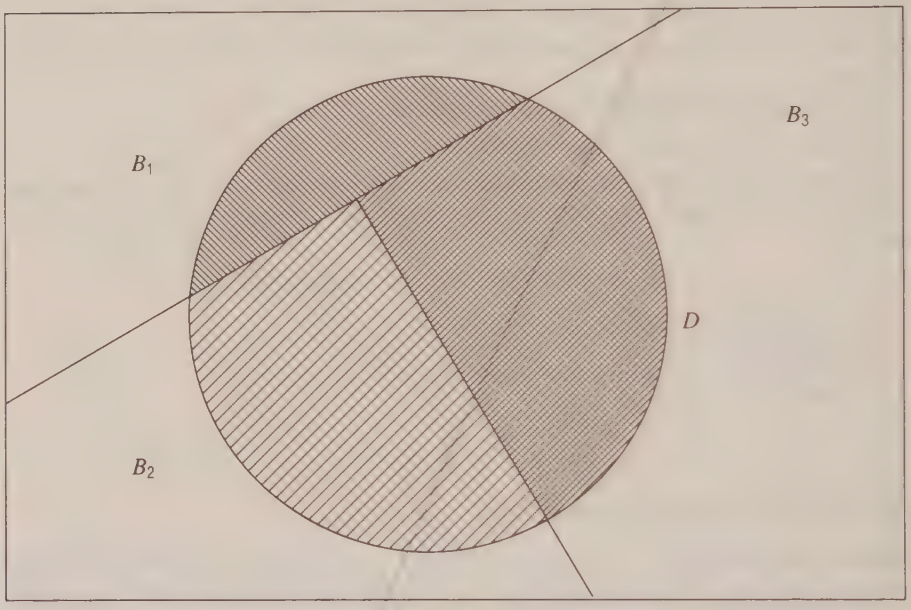


Figure 9.7 Venn diagram representation of the law of total probability

EXAMPLE 9.7

The police detective division of a certain city uses a lie detector to question criminal suspects. According to statistics collected over a course of time by the division, any suspect will fall into one of three groups, namely:

B_1: suspect is a habitual criminal
B_2: suspect is an occasional criminal
B_3: suspect has no criminal record

The relative frequencies of the groups are known, namely:

$$\frac{f_1}{n} = 0.15$$

$$\frac{f_2}{n} = 0.65$$

$$\frac{f_3}{n} = 0.20$$

Let D = accused has committed the crime involved, and is caught out by the lie detector.

Experience has shown that when an accused is questioned about a particular crime, the probability that he will be exposed by the lie detector if he has committed the crime is different for the three groups, namely:

$P(D \mid B_1) = 0.60$

$P(D \mid B_2) = 0.75$

$P(D \mid B_3) = 0.90$

Suppose a suspect who is totally unknown to the police is tested. What is the probability that he committed the crime concerned and that he will be caught out by the lie detector?

According to the law of total probability,

P(accused has committed the crime and is caught out)

$= P(D)$

$= P(B_1)\,P(D \mid B_1) + P(B_2)\,P(D \mid B_2) + P(B_3)\,P(D \mid B_3)$

$= (0.15 \times 0.60) + (0.65 \times 0.75) + (0.20 \times 0.90)$

$= 0.7575$

≈ 0.76

3.8 Bayes' rule

By applying the above rules of probability a variety of more advanced rules of probability can be derived. One of the most important, namely Bayes' rule, will suffice here.

Take events $B_1, B_2, \ldots, B_m$, which are all mutually exclusive and for which $B_1 \cup B_2 \cup \ldots \cup B_m = S$, the certain event, applies. Let D be any other event based on the same sample space S. According to **Bayes' rule**, for $P(D) > 0$ and $P(B_j) > 0, j = 1, 2, \ldots, m$,

$$P(B_i \mid D) = \frac{P(B_i)\,P(D \mid B_i)}{\sum P(B_j)\,P(D \mid B_j)}$$

holds.

Bayes' rule has several applications, including the following two:

(a) If an event D has occurred due to one of m possible 'causes' $B_1, B_2, \ldots, B_m$, the probability that B_i was the 'cause' for D having taken place, can be determined from this rule.

(b) Take a population of individuals that can be divided into m groups $B_1, B_2, \ldots, B_m$. The probability of a randomly selected individual being a member of the jth group is given by $P(B_j)$, and the probability of individuals from group B_j possessing a certain characteristic D, is given by $P(D \mid B_j), j = 1, 2, \ldots, m$. If the randomly selected individual's group is not known, but it is clear that he possesses characteristic D, the probability of his being a member of a specified group B_i can be updated by calculating $P(B_i \mid D)$.

EXAMPLE 9.8

All radios manufactured by a certain factory are tested by one of three inspectors, Benade, Botha and Brink. Due to the extent of the work the testing cannot be done on a one hundred per cent basis. However experience has shown that only 2%, 1% and 4% of the radios approved by Benade, Botha and Brink respectively have some or other defect. The performance of the three inspectors also differs so much that 40%, 35% and 25% of the total production is tested by the three inspectors respectively. If a newly sold radio is sent back to the factory because of a defect, what are the respective probabilities of it having been tested by these three inspectors?

Indicate by means of B_1, B_2 and B_3 the events that the defective radio was tested by Benade, Botha and Brink respectively, and by D the event that a defective radio is sold. It is clear that one of the three inspectors was the cause of the defective radio slipping through, so B_1, B_2 or B_3 were the three possible 'causes' of D realizing. The probabilities of the events B_1, B_2 and B_3 are known, namely

$$P(B_1) = 0.40;\ P(B_2) = 0.35 \text{ and } P(B_3) = 0.25$$

Furthermore, the conditional probability of a radio being defective, given that it was tested by a specific inspector, is also known, namely

$$P(D|B_1) = 0.02;\ P(D|B_2) = 0.01 \text{ and } P(D|B_3) = 0.04$$

In accordance with Bayes' rule, the probability that the defective radio was tested by inspector Benade is

$$\begin{aligned} P(B_1|D) &= \frac{P(B_1)\ P(D|B_1)}{\sum P(B_j)\ P(D|B_j)} \\ &= \frac{(0.40 \times 0.02)}{(0.40 \times 0.02) + (0.35 \times 0.01) + (0.25 \times 0.04)} \\ &= \frac{0.008}{0.0215} \\ &\approx 0.372 \end{aligned}$$

Similarly, it follows that

$$\begin{aligned} P(B_2|D) &= \frac{P(B_2)\ P(D|B_2)}{\sum P(B_j)\ P(D|B_j)} \\ &= \frac{(0.35 \times 0.01)}{(0.40 \times 0.02) + (0.35 \times 0.01) + (0.25 \times 0.04)} \\ &= \frac{0.0035}{0.0215} \\ &\approx 0.163 \end{aligned}$$

$$\begin{aligned} P(B_3|D) &= \frac{P(B_3)\ P(D|B_3)}{\sum P(B_j)\ P(D|B_j)} \\ &= \frac{(0.25 \times 0.04)}{(0.40 \times 0.02) + (0.35 \times 0.01) + (0.25 \times 0.04)} \end{aligned}$$

$$= \frac{0.0100}{0.0215}$$

$$\approx 0.465$$

The most likely person to have 'accepted' the defective radio is therefore inspector Brink.

EXAMPLE 9.9

An insurance company approaches an enterprise with many employees with a view to short-term group insurance. Experience has shown the insurance company that enterprises such as this one can be divided into one of three categories, namely:

B_1: low-risk group
B_2: medium-risk group
B_3: high-risk group

The insurance company bases its classification of enterprises on the probability that the total claim of an enterprise would be more than an amount K.

Let D = enterprise's total claim more than an amount K

Furthermore, reports by insurance companies show that the distribution of this type of enterprise across the three categories (expressed as probabilities) is as follows:

$P(B_1) = 0.25$
$P(B_2) = 0.60$
$P(B_3) = 0.15$

The probabilities of a specific group submitting a total claim of more than K, are given respectively by

$P(D \mid B_1) = 0.01$
$P(D \mid B_2) = 0.08$
$P(D \mid B_3) = 0.60$

The enterprise concerned claims to be a member of the low-risk group and the insurance company and the board of directors of the enterprise agree that the board will pay the premium of the B_1 group (which is less than the premium of the other two groups) during the first year of their commitment. At the end of that year the enterprise will be reclassified if necessary.

At the end of the first year the total claim for the enterprise indeed turns out to have exceeded K. Determine the probabilities of the enterprise being a member of the various groups.

The denominator of the Bayes expression is

$$\sum P(B_i)\, P(D \mid B_i) = (0.25 \times 0.01) + (0.60 \times 0.08) + (0.15 \times 0.60)$$
$$= 0.1405$$

It therefore follows from Bayes' rule that

$$P(B_1 \mid D) = \frac{P(B_1)\ P(D \mid B_1)}{\sum P(B_i)\ P(D \mid B_i)}$$

$$= \frac{0.25 \times 0.01}{0.1405}$$

$$\approx 0.018$$

$$P(B_2 \mid D) = \frac{0.60 \times 0.08}{0.1405}$$
$$\approx 0.342$$
$$P(B_3 \mid D) \approx 0.641$$

The probability of the enterprise being a member of the low-risk group thus decreased once their claim pattern became known. In other words: At the beginning of the negotiations the probability of the enterprise being a member of the high-risk group was 0.15, while after one year it appears to be 0.641.

4 Counting techniques

Counting techniques play an important role in determining probabilities. When applying the classical definition of probability it is important, for example, to count both the number of equally possible outcomes of a stochastic experiment and the number of outcomes that favour the realization of the event being studied. Some of the most important counting techniques will now be discussed.

4.1 The basic counting rule

When analysing a stochastic experiment, it is often possible to distinguish more than one stage. The drawing of a single random number can be regarded as a **one-stage experiment**, the subsequent drawing of two random numbers as a **two-stage experiment**, et cetera. Generally we refer to a **multistage experiment** if the experiment involves the execution of multiple actions such as the repeated drawing of numbers or the repeated rolling of a dice. The following rule forms the basis for counting techniques, or **combinatorial analysis**.

> If one action in a k-stage experiment can be carried out in m_1 ways, a second action in m_2 ways, ..., the kth action in m_k ways, the entire experiment can be carried out in a total of $m_1 \times m_2 \times \ldots \times m_k$ ways. This rule is known as the **basic counting rule**.

EXAMPLE 9.10

A courier service receives parcels at two collection points. A fleet of three vehicles is used to deliver the parcels at four points. An important parcel is lost after having been received at one of the collection points. How many possibilities have to be checked, if:

(a) it is known that the parcel was received at collection point 1;

(b) it is not known at which collection point the parcel was received?

According to the basic counting rule, it now follows that

(a) Number of possibilities $= 3 \times 4$
$= 12$

(b) Number of possibilities $= 2 \times 3 \times 4$
$= 24$

A visual aid that usually provides insight into problems with counting techniques is the **tree diagram** described in Section 3.5. A tree diagram of the various possibilities in case (b) of Example 9.10 is given in Figure 9.8.

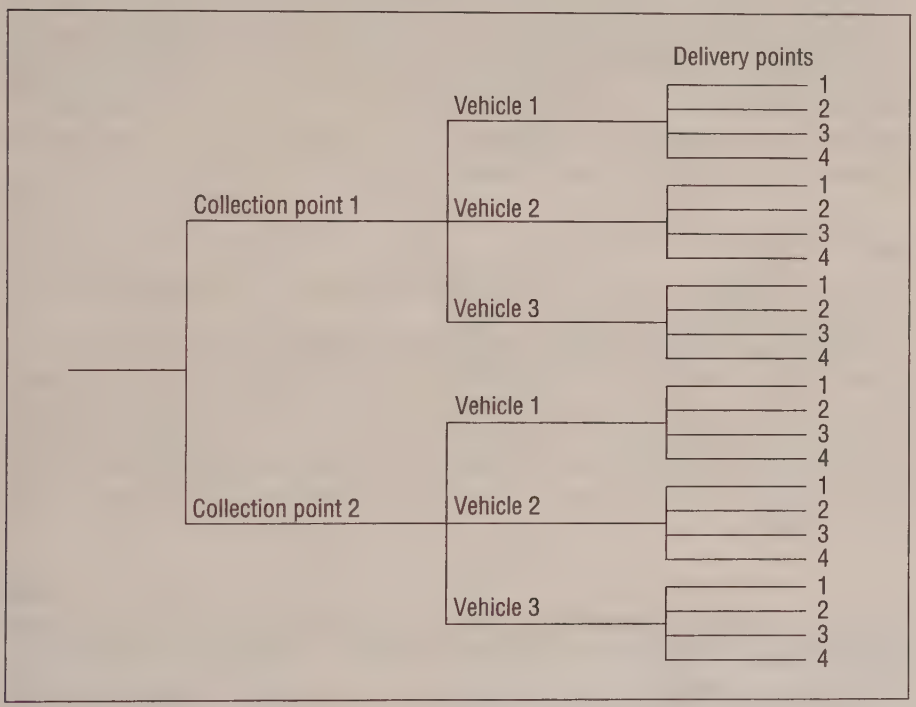


Figure 9.8 Tree diagram of the various possibilities of receiving, transporting and delivering a parcel

The tree diagram in Figure 9.8 is a schematic representation of the total number of possibilities to be explored in the search for the missing parcel. In many cases it is difficult and even practically impossible to draw a tree diagram of a multistage experiment and thus physically count the outcomes. However, it is not always necessary to count all cases physically, since the counting process can be carried out by means of different algebraic formulas. These formulas are expressed in terms of so-called **permutations** and **combinations**.

4.2 Permutations

In word-building competitions competitors sometimes want to know how many different words can be formed by a group of letters. Various factors play a role here, namely:

(a) May the letters be repeated?
(b) How many letters may be used per word?

This type of problem can be solved by using **permutations**. Study the following definition in terms of four symbols (extension of this to any number of symbols is direct):

> All possible arrangements of the four objects a, b, c, d, namely $abcd$, $abdc$, ..., $dcba$, are known as the **permutations** of the objects.
>
> There are $n! = n \times (n-1) \times (n-2) \times \ldots 3 \times 2 \times 1$ different possible permutations of the n objects $v_1, v_2, \ldots, v_n$.

Note that $n!$ is pronounced: 'n factorial'. The argument in the following example illustrates the derivation of this formula.

EXAMPLE 9.11

How many possible four-letter 'words' can be formed from the four letters a, b, e and g if each letter may be used only once?

The first position can be filled by one of the four letters, namely a, b, e or g. The second position can therefore be filled by any one of the remaining three letters (e.g. by a, b or e if the first was g). Similarly, the third position can be filled by any of the two remaining letters, the fourth letter then taking the remaining position. According to the basic counting rule, the four positions can therefore be filled in $4 \times 3 \times 2 \times 1 = 4! = 24$ ways, that is, 24 different possible four-letter words can be formed.

The above example immediately raises the question: How many different 'words' can be formed from the four letters a, b, e, g if each of the letters may be used more than once?

The following argument now applies: The first position in the word can be filled by any of four possible letters, the second by any of four possible letters, ..., and the fourth by any of four possible letters. In total the four positions can therefore be filled in $4 \times 4 \times 4 \times 4 = 4^4 = 256$ different ways.

Since each letter can now be used more than once, the term 'permutation' can create a problem here. We shall therefore use a more general term, namely **fourfold** for the letter grouping $aabc$, or in general, m-fold for the objects $v_1 v_2 v_3 \ldots v_m$.

> There are m^m **different possible m-folds** that can be formed from the objects $v_1, v_2, \ldots, v_m$ if each object may appear **any number of times** in an m-fold.

Let us look at a more general situation in which four-letter words have to be formed from ten letters. Take in general n objects $v_1, v_2, \ldots, v_n$ and let $m < n$.

There are

$$n!^m = \frac{n!}{(n-m)!} \qquad \text{(pronounced as '}n \text{ permutation } m\text{')}$$

different possible permutations of m objects that can be formed from the n objects $v_1, v_2, \ldots, v_n$ if each object may appear only once in the permutation.

There are n^m different possible m-folds that can be formed from the n objects $v_1, v_2, \ldots, v_n$ if each object may appear any number of times in an m-fold.

Both formulas follow directly from the application of the basic counting rule. Thus, for example, in the first case:

The first position in the permutation can be filled by any of the n objects, the second by any of the remaining $(n - 1)$ objects, ..., and the mth position by any of the remaining $(n - m + 1)$ objects. From the basic counting rule it therefore follows that:

Number for permutations

$$= n \times (n-1) \times \ldots \times (n-m+1)$$

$$= \frac{n \times (n-1) \times \ldots \times (n-m+1) \times (n-m) \times \ldots \times 3 \times 2 \times 1}{(n-m) \times \ldots \times 3 \times 2 \times 1}$$

$$= \frac{n!}{(n-m)!}$$

EXAMPLE 9.12

How many different four-letter words can be formed from the 10 letters $a, b, \ldots, j$ if each letter may be used

(a) only once;
(b) any number of times?

(a) number of four-letter words $= 10!^4$

$$= \frac{10!}{(10-4)!}$$

$$= \frac{10!}{6!}$$

$$= 10 \times 9 \times 8 \times 7$$

$$= 5\,040$$

(b) number of four-letter words $= 10^4$

$$= 10\,000$$

4.3 Combinations

One of the most important applications of sampling in the manufacturing sector is **acceptance sampling**. According to this application a random sample of m items is drawn from a batch of n items and the m items are then tested for defects. Based on the number of defects identified among the m items in the sample, the purchaser can decide whether or not to accept the entire batch. A logical question that could arise is: How many different samples of size m can be drawn from n items?

For the sake of simplicity, take a container with ten numbered balls (numbers 0, 1, 2, ..., 9) that are identical in all respects except for the numbers. Four balls can be drawn simultaneously from the container, which was well shaken beforehand. The result can, for example, be represented as follows: {4, 3, 9, 6}. It is clear that only the **content** of the sample is important, and that there is no mention of an **order** or permutation with regard to the four numbers. Such a grouping of numbers is known as a **combination**. We can represent all possible combinations of four items from ten different items as follows: {0, 1, 2, 3}, {0, 1, 2, 4}, ..., {6, 7, 8, 9}. The question is now: How many possible combinations are there? We can use the following argument:

Suppose there are k different combinations of size 4 from ten different items. For each combination of 4 numbers, 4! different permutations can be formed. There are therefore $k \times 4!$ different permutations. However, we know from the above treatment of permutations that $10^{!4}$ different permutations can be formed from the 10 numbers 0, 1, 2, ..., 9. These two expressions should therefore render the same result, i.e.

$$k \times 4! = 10^{!4}$$

From this it follows that

$$\begin{aligned} k &= \frac{10^{!4}}{4!} \\ &= \frac{10!}{(10-4)!\,4!} \\ &= 210, \text{ after simplification.} \end{aligned}$$

It is customary to designate the total number of combinations of size 4 from the ten objects $v_1, v_2, \ldots, v_{10}$ by means of the notation $\binom{10}{4}$ (pronounced as '10 combination 4').

From the above discussion the following general result can now easily be understood:

There are

$$\binom{n}{m} = \frac{n!}{(n-m)!\,m!}$$

different possible combinations of size m from the n objects $v_1, v_2, \ldots, v_n$.

Notes

(a) The following expressions are important in the application of permutations and combinations:

(i) $n! = n \times (n-1)!$
$= n \times (n-1) \times (n-2)!$, etc

(ii) $0! = 1$

(iii) $\binom{n}{0} = 1$

(iv) $\binom{n}{m} = \binom{n}{n-m}$

(b) It is clear that $n!$ increases rapidly with an increase in n. Thus, for example, 10! is more than 3.6 million; and 11! is almost 40 million.

There are several tables from which $n!$ and $\binom{n}{m}$ can be found for different values of m and n, while many pocket calculators contain $n!$ as a standard function. However, please note that the value for $n!$ obtained with the aid of a pocket calculator is only accurate to the number of (first) significant digits that the calculator can indicate.

When simplifying an expression that contains the factorials of large numbers, it is advisable to express the factorials as products and to cancel, where possible, the factors in the numerator and denominator of the expression, for example:

$$\binom{40}{10} = \frac{40!}{10!\,30!}$$

$$= \frac{40 \times 39 \times \ldots \times 31 \times 30!}{10!\,30!}$$

$$= \frac{40 \times 39 \times \ldots \times 31}{10 \times 9 \times \ldots \times 1}$$

$$= 847\,660\,528, \text{ after simplification.}$$

EXAMPLE 9.13

The freight on a train with 40 trucks of sunflower seed from a certain agricultural region has to be graded according to the percentage of oil in the seed. If the seed contains more than a prescribed percentage of oil, it is regarded as highly suitable for making products such as cooking oil and margarine (high-grade seed). At the grain silo where the seed is received, the content of only five trucks can be tested by means of sampling. It is assumed that the sampling process from a truck is such that the result is completely reliable, in other words, if the test sample indicates a certain quality of seed, it can be accepted that the content of the entire truck is of

like quality. If there are twelve trucks with a percentage of oil lower than the prescribed level, what is the probability of the sample of five trucks containing

(a) exactly two trucks
(b) at most two trucks

with lower grade seed?

Application of the combination formula, the basic counting rule and the classical definition of probability reveals the following:

Number of possible samples of size 5 from a total of 40 trucks: $\binom{40}{5}$

Number of samples of size 5 of which two from twelve trucks with lower grade seed and three from 28 trucks with higher grade seed: $\binom{12}{2}\binom{28}{3}$

$$P(\text{exactly 2 trucks with lower grade seed}) = \frac{\binom{12}{2}\binom{28}{3}}{\binom{40}{5}}$$

$$= \frac{66 \times 3\,276}{658\,008}$$

$$= 0.32859$$

$$\approx 0.33$$

This answers the first question.

Similarly, to answer the second question it follows that

$$P(\text{at most 2 trucks with lower grade seed})$$

$$= \frac{\binom{12}{0}\binom{28}{5}}{\binom{40}{5}} + \frac{\binom{12}{1}\binom{28}{4}}{\binom{40}{5}} + \frac{\binom{12}{2}\binom{28}{3}}{\binom{40}{5}}$$

$$= \frac{(1 \times 98\,280) + (12 \times 20\,475) + (66 \times 3\,276)}{658\,008}$$

$$= 0.85135$$

$$\approx 0.85$$

4.4 Methods of approximation

Because $n!$ increases so rapidly, it is often impractical or cumbersome to determine exactly expressions like $\binom{100}{30}$. The question arises, for instance, how we can solve the problem of Example 9.13 if the train had 160 trucks, the content of 48 trucks was of a lower grade, and a sample from twenty trucks had to be tested.

This problem can be solved by means of an approximation method. Study the following simulation of the problem:

A container has N identical items, of which a ratio of p has a certain characteristic, say characteristic D (which, for instance, can be defective items in a batch of new items). A random sample of size n items is drawn from the container. The probability of x of the items in the sample having characteristic D is

$$P(x \text{ items having characteristic } D) = \frac{\binom{Np}{x}\binom{N(1-p)}{n-x}}{\binom{N}{n}}$$

If N is very large, n much smaller than N and p is not too close to 0 or 1, it can be proven that this probability, if approximated, is equal to $\binom{n}{x}p^x(1-p)^{n-x}$, in other words

$$P(x \text{ items having characteristic } D) = \frac{\binom{Np}{x}\binom{N(1-p)}{n-x}}{\binom{N}{n}}$$

$$\approx \binom{n}{x}p^x(1-p)^{n-x}, \text{ provided } n << N \text{ and } p \text{ is not too close to 0 or 1.}$$

EXAMPLE 9.14

We again take the problem set out above, with samples of sunflower seed being drawn from a train with a large number of trucks. Suppose there are 160 trucks and 48 of them contain lower grade seed. A sample of twenty trucks is drawn. To summarize, the following applies:

$N = 160$

$p = \frac{48}{160} = 0.3$

$n = 20$

$$\text{Then } P(8 \text{ trucks with lower grade seed}) = \frac{\binom{48}{8}\binom{112}{12}}{\binom{160}{20}}$$

$$\approx \binom{20}{8}(0.3)^8(0.7)^{12}$$

$$= 0.11440$$

$$\approx 0.11$$

Note

The last two examples are applications of the so-called **hypergeometric** and **binomial theory** that will be discussed further in Chapter 10.

Exercises

1 A stochastic experiment comprises the random selection of one house per population of houses in the four cities given in Appendix A, Data Set 1. Define the events A, B and C as follows:

A: house with small plot (size 1);
B: house with medium plot (size 2); and
C: house with large plot (size 3).

(a) Use the data of the first 100 houses of each of the four cities to draw up a two-way frequency table of city × stand size.
(b) Use the frequency limit definition to find approximate values for $P(A)$, $P(B)$ and $P(C)$ for each of the four cities.

2 If a two-way frequency distribution of city by plot size is prepared for the entire Data Set 1 of Appendix A, the table below is obtained. If the data set is regarded as a population and the same stochastic experiment as in Exercise 1 is carried out, use the classic definition to determine the probabilities $P(A)$, $P(B)$ and $P(C)$ for each of the four cities. Compare these probabilities with the approximate probabilities of Exercise 1.

Two-way table of city by plot size for Appendix A, Data Set 1

City	Plot size		
	Small	Medium	Large
Cape Town	282	118	6
Durban	100	174	102
Johannesburg	387	373	83
Pretoria	53	341	40

3 A container holds twelve identical cards that differ only in colour: six are red, four are yellow and two are white. One card is randomly drawn from the container. Consider the following events:

A: red card; and
B: yellow card.

Using the classic definition determine the following probabilities:
(a) $P(A)$, $P(B)$, $P(A \cap B)$ and $P(A' \cup B')$
(b) $P(\overline{A \cup B})$
(c) P (white card). Explain the connection between this answer and the answer to (b).

4 A stochastic experiment comprises the random selection of one house from the 2 059 houses in Appendix A, Data Set 1. Again define the events A, B and C as in Exercise 1, and let:

K: sampled house is a Cape Town house;
D: sampled house is a Durban house; and
J: sampled house is a Johannesburg house.

Using the two-way table in Question 2, determine the probability of the following events:

(a) $P(A \cap K)$
(b) $P(A \mid K)$
(c) $P(J \cup C)$
(d) $P(\overline{K \cup D \cup J})$
(e) $P(K \cap \overline{A \cup B})$
(f) $P(C \mid J \cup D)$

5 Consider the events A, B and C where $P(A) = \frac{1}{4}$, $P(B) = \frac{1}{2}$ and $P(A \cup B) = \frac{3}{5}$. Determine:

(a) $P(A \cap B)$
(b) $P(\overline{A} \cap B)$
(c) $P(\overline{A} \cap \overline{B})$
(d) $P\{(A \cap \overline{B}) \cup (\overline{A} \cap B)\}$

(Clue: draw a Venn diagram).

6 If an experiment is attempted only once, only one out of three probable results is possible, A, B or C, with probabilities $P(A)$, $P(B) = 2P(A)$ and $P(C) = 5P(A)$. Determine:

(a) $P(A)$
(b) $P(A \cup B)$

7 In a stochastic experiment there are, among others, two events, A and B with $P(A) = \frac{2}{5}$ en $P(B) = \frac{1}{2}$. Determine $P(A \cup B)$ if A and B are

(a) mutually exclusive; and
(b) mutually independent.

8 In a survey conducted among students the following observations were made regarding their music preferences:
50 enjoy heavy classical music;
60 enjoy light classical music;
72 enjoy light pop music;
30 enjoy heavy classical and light pop music;
24 enjoy light classical and heavy classical music;
18 enjoy light pop and light classical music; and
12 enjoy light pop and any classical music.

(a) Draw a Venn diagram to depict the different groups. How many students are involved in the survey?
(b) What percentage of students enjoy
    (i) heavy classical music;
    (ii) pop music;
    (iii) any light music; and
    (iv) classical music?
(c) One student is chosen at random. What is the probability that this student enjoys light classical music, given that he/she does not enjoy heavy classical music?

9 Amongst others, events A and B are defined for the purpose of a stochastic experiment. Use a Venn diagram to find an expression in terms of $P(A)$, $P(B)$ and $P(A \cap B)$ for $P\{(A \cap \overline{B}) \cup (\overline{A} \cap B)\}$.

10 A dice is loaded so that the probability of a result is directly proportional to the result. The dice is rolled once. Let A: a five or a six in one roll. Determine:
(a) the probabilities for the various possible results;
(b) the probability of an even number;
(c) which probability is the largest, $P(A)$ or $P(\overline{A})$?

11 A black dice and a red one are rolled simultaneously. Let:
A: the results on both dice are the same;
B: the value of the red dice is greater than, or equal to, the value of the black dice.
(a) Express the sample space and events A and B in terms of set theory.
(b) Determine $P(A)$, $P(B)$ and $P(A \mid B)$.

12 Prove the following expression for two events E and F:

$$P(E \mid \overline{F}) = \frac{P(E) - P(E \cap F)}{1 - P(F)}$$

13 Two men and three women apply for two identical jobs. Should the employer choose to appoint two people at random, express the following events in set theory:
S: the certain event;
A: one man and one woman appointed;
B: at least one woman appointed.
Determine $P(A \cup B)$, $P(A \cap B)$ and $P(\overline{A})$.

14 Determine the values of:
(a) 8! and 9!
(b) $8^{!4}$ and $8^{!5}$
(c) $\binom{8}{5}$ and $\binom{8}{3}$
(d) $\dfrac{\binom{10}{4}\binom{10}{6}}{\binom{20}{10}}$
(e) x if $\dfrac{\binom{6}{x}}{(x+1)} = \dfrac{6^{!x}}{24}$

15 Determine the values of
(a) $\binom{6}{0}, \binom{6}{1}, \ldots, \binom{6}{6}$
(b) $\sum_{r=0}^{6} \binom{6}{r}$ and 2^6. What do you notice?
(c) n if $\sum_{r=0}^{n} \binom{n}{r} = 4\,096$

16 A motor vehicle licensing department issues registration numbers compiled as follows:
3 letters of the alphabet (excluding A, E, I, O, U and Q)
3 numbers
the letter T.
(a) How many registration numbers can be compiled?
(b) If the letter T at the end could also be any one of the remaining 25 letters of the alphabet, how many registration numbers could then be compiled?

17 Twenty-four new electronic components are in a container with sixteen used components of which four are defective. The two sets of components accidentally get mixed up in the container. Four components are randomly taken from the container. Determine the probability that:
(a) all four are defective;
(b) exactly two are defective;
(c) two defective components (D) and two faultless components (N) will be drawn in the following order: $D\ N\ D\ N$.

18 There are balls of three colours in a container: twenty red balls, fifteen black balls and twelve yellow balls. Ten balls are chosen at random. Determine the probability of:
(a) five red, three black and two yellow balls;
(b) five red and five black balls; and
(c) five red balls
being drawn.

19 In an agricultural experiment, three soil types, four fertilizer mixtures and two grain variants need to be studied to determine which combination will yield the best crop. Each combination is tested on a half-hectare plot. How many of these plots will be needed if each combination
(a) is tested on only one plot?
(b) is tested on five plots?
(c) Supply answers to the above questions if each combination is also subjected to tests of two types of irrigation.

20 A selection committee comprising three men and two women has to decide which of two equally competent candidates, one male and one female, should be appointed to a position. The male selectors all choose the man and the two female selectors choose the woman. A dispute ensues, whereupon it is decided that the probability of the outcome be determined under the assumption that the choice of each of the selectors is completely random.
(a) In how many ways can the committee be divided if three selectors choose the man and two selectors choose the woman?
(b) What is the probability that the outcome given above was random?

21 A card is randomly drawn from a container with twelve cards numbered 1, 2, ..., 12. Let

A: the number smaller than 5
B: the number between 3 and 8 (both included).
(a) Determine $P(A)$, $P(B)$, $P(A \cap B)$ and $P(A)\,P(B)$.
(b) Are A and B mutually independent? Motivate.

22 From a survey conducted at a service station, it becomes clear that 5% of the clients are dissatisfied with repair jobs on their vehicles. Fifty per cent of the complaints received are from clients whose repairs are done by mechanic A, yet it is known that A does 60% of all repairs. Given that A does the repair job, determine the probability that:
(a) a client will be dissatisfied with his car; and
(b) a client will be satisfied with his car.

23 Suppose A and B are mutually exclusive with $P(A) > 0$ and $P(B) > 0$. Prove that A and B are dependent.

24 A dealer in car batteries mistakenly allows two defective and eight faultless batteries to become mixed up. He systematically tests all the batteries but stops once he has found the second defective one. Determine the probability that:
(a) the second defective battery will be found during the fourth test; and
(b) the second defective battery will be found during the last test.
(c) Determine the probability that the second defective battery will be found at or before the fourth test if the first defective battery is found during the first two tests.

25 A student answers a multiple test examination paper. Each question has a total of five possible answers of which only one is correct. Obviously he will make the correct choice for the questions for which he knows the answer. If he is unsure, he chooses the answer about which he feels most certain, but if he has no idea of any possible answer and has to guess, he makes a random choice. If, for questions he is unsure of, his answers are 60% correct, then what percentage of all correct answers is mere guesswork? Assume that for an equal number of questions he knows the answer, he is uncertain or does not know the answer at all.

26 Based on his experience, an insurance agent has classified all his clients as belonging to one of the following groups:
A: "I am not interested in any new policy."
B: "I may possibly be interested – tell me more."
C: "I am definitely interested in some or other form of investment or insurance."
According to his records 40% of his clients fall under category A, 35% under B and 25% under C. The probability of a client from category B buying a policy is 20% and for a client from category C, 30%. Should this agent see a new client, what is
(a) the probability that he will sell a policy?

(b) the probability that if the client buys a policy he will fall under category *C*?

27 There are ten cards numbered 0, 1, 2, ..., 9 in a container. A stochastic experiment comprises the random withdrawal (without replacement) of five cards from the container. How many numbers can be derived from the digits of the five cards if:
(a) the five digits are written down in the order that the cards are drawn;
(b) the five digits are written down in ascending order;
(c) the compiled five-digit number has to be an odd number.

28 A motorist who always travels to work along the same route has noticed that at a certain traffic light the green phase lasts 30 seconds, the red phase lasts 17 seconds and the yellow phase 3 seconds. If he assumes that his time of arrival at the intersection is random, determine the probability that:
(a) on five consecutive days he will find the light to be green;
(b) on five days out of ten he will find the light to be green; and
(c) on five days he will find the light to be either yellow or red.

29 The so-called 'game of matches' is played with two identical packs of playing cards numbered from 1 to 10. After each pack has been well shuffled both packs are placed face down. The cards from each pack are then simultaneously turned over one by one from the top down. If, at any given point in the game, two cards with the same number are turned over, a 'match' is said to have occurred.
Determine the probability of a match occurring:
(a) with the first attempt;
(b) with both the first and last attempts;
(c) with all ten attempts; and
(d) with at least one attempt.

30 A shoe salesman has shown a client ten pairs of shoes and has neglected to pack each pair back into its own box after the client has tried them on. He then puts the shoes back into the boxes without checking. What is the probability that:
(a) the first pair of shoes is put into the right box;
(b) the first and the last pair of shoes are put into the right boxes; and
(c) all ten pairs of shoes have been put into the right boxes?
(Clue: Game of matches: see Exercise 29).

31 There are three colours of beads in a container: red, blue and white. One colour comprises half the beads while the remaining half is comprised of equal numbers of each of the other two colours. You must determine the composition of the beads in the container by drawing three beads consecutively and replacing each bead after noting its colour and before drawing another. The order in which the three beads are drawn is red, blue and red.

Determine the ratio after
(a) the first bead;
(b) the second bead; and
(c) the third bead
has been drawn.

32 Computer monitors are packed into two crates and sent from Taiwan to South Africa. Each crate holds twenty monitors. Suppose crate 1 contains five defective monitors and crate 2 only one defective monitor. One crate is chosen randomly and from this crate one monitor is chosen as a random sample for testing. The monitor proves to be defective. What is the probability that the defective monitor was packed in crate 1?

33 Suppose 5% of the population of a certain city has a certain disease. In 90% of the cases the blood test that is available will correctly identify that a person has the disease. Unfortunately, the same test also incorrectly gives a positive test result in 15% of cases where there is no infection present. One person is selected randomly from the population group and the test is positive. What is the probability that this person does have the disease?

34 In a stochastic experiment a coin is tossed until heads comes up. Thus 1 or 2 or 3 or ... trials may be needed to complete the experiment.
(a) Express the sample space S of the experiment.
(b) Prove that $P(S) = 1$.
(c) Determine the probability that no more than five trials will be required to complete the experiment.
(d) Determine a general expression for the probability that the experiment can be completed in n trials. What must the value of n be if $P(\text{experiment completed in } n \text{ trials}) \geq 0.99\%$?

CHAPTER 10

Probability models

1 Introduction

In Chapter 3 we showed how a frequency distribution for a raw data set can be drawn up and represented graphically by a histogram or bar chart. In Chapters 4 and 5 we showed how conclusions regarding the general shape of the frequency distribution can be drawn from descriptive measures. The question is: what will such a frequency distribution and its graphic representation look like if the total population were to be studied with respect to the characteristic concerned? A full investigation or census is seldom carried out, since the population is usually extensive and, for various reasons, it is impractical to study the entire population (see Chapter 2, Section 2). An example of such a population is the 2 000 residences of which several characteristics are given in Appendix A.

However, the concept **population** is often also used in a far wider sense, for example, to refer to the population of outcomes when a dice is repeatedly rolled or to the population of all the light bulbs ever manufactured or still to be manufactured by a factory. Clearly, the values of the various population characteristics (for example: house prices or outcomes in rolling a dice) do have to have some sort of distribution. This distribution is called a **theoretical distribution**, as opposed to the frequency distribution or **empirical distribution** of a sample. Each measurement of the population characteristic is regarded as the observation of a variable, for example, the price of the house or the outcome in rolling a dice. Such a variable is known as a **stochastic variable**.

The next section deals with certain basic concepts that have a bearing on stochastic variables, while a number of theoretical distributions are dealt with in the remaining sections of this chapter.

2 Basic concepts

2.1 Stochastic variables and associated distributions

Consider the following two examples:

(a) An electrical appliance retailer sends all defective items returned by clients to a firm for repairs. The firm doing the repairs visits the retailer once a week to collect defective items and return repaired ones. The two parties

have agreed that all items will be returned to the retailer within six weeks, so that a client has to wait between one and six weeks for his repaired item to be returned. If a new client hands in an item for repair, the time he/she has to wait can be regarded as the execution of a stochastic experiment. The stochastic variable associated with this is

X = number of weeks of waiting

and the waiting times of ten randomly selected clients can be the following:

3, 1, 4, 6, 2, 2, 4, 3, 1, 5.

In this case we can say that the stochastic variable X has consecutively **taken on** the ten values.

A **stochastic variable** allocates a real value to each possible outcome of a stochastic experiment.

In order to distinguish between a stochastic variable and a certain value of that variable, we shall use the following notation:

A stochastic variable is designated by a **capital letter** and specific values of this variable by **lower-case** letters.

Thus, for example, the stochastic variable 'number of weeks waiting time' above is indicated by X, and a random sample of ten values of X can generally be indicated by $x_1, x_2, \ldots, x_{10}$.

(b) As the second example we look at the experiment in which the price per square metre is determined in a random sample of five residences in the Johannesburg area. Suppose the associated stochastic variable

Y = price per square metre

takes on the following five values:

710, 1 250, 906, 864, 437 rand.

It is clear that the stochastic variables 'number of weeks' waiting time' and 'price per square metre' differ. The first variable, X, can take on only fixed, isolated values and is known as a **discrete stochastic variable**, while Y can take on any meaningful value within a certain interval of values and is known as a **continuous stochastic variable**. Note that, for practical reasons, the price per square metre, Y, is given in round numbers. Theoretically, however, Y can be given to, for example, the nearest cent.

The set of all possible values of a stochastic variable is known as the **domain** of the stochastic variable.

Since the domain of a stochastic variable is always a set of real numbers, we designate the domain of X by R_X, that of Y by R_Y, and so forth. The domain of X, the number of weeks' waiting time, is

$$R_X = \{1, 2, 3, 4, 5, 6\},$$

while the domain of Y, the price per square metre, is

$$R_Y = \{y \mid a \leq y \leq b\}$$

with a and b the minimum and the maximum price per square metre for the purchase of houses in the Johannesburg area.

EXAMPLE 10.1

Refer to Example (a) at the beginning of this section and consider a random sample of the waiting times of 120 customers as indicated in the frequency distribution given in Table 10.1.

Table 10.1 Frequency distribution of the waiting times (number of weeks) of 120 customers for repairs to electrical appliances

Number of weeks x	Frequency f_x	Relative frequency $f_x / 120$
1	9	0.075
2	18	0.150
3	27	0.225
4	34	0.283
5	21	0.175
6	11	0.092
	120	1.000

According to the frequency limit definition of probability (see Chapter 9, Section 2) and Table 10.1, the probability of a randomly selected customer waiting exactly three weeks can be set at approximately $\frac{27}{120} = 0.225$. A more accurate value of this probability, and of the probabilities of having longer and shorter waiting periods, can be determined by observing a larger sample of clients. However, we find that the approximate frequency distribution (or **empirical distribution** of X) is symmetrical with $\frac{f_1}{120}$ and $\frac{f_6}{120}$ of the same order of magnitude, $\frac{f_2}{120}$ and $\frac{f_5}{120}$ of the same order of magnitude, and $\frac{f_3}{120}$ and $\frac{f_4}{120}$ of the same order of magnitude. Furthermore, $f_1 : f_2 : f_3 \approx 1:2:3$. If an exceptionally large sample of customers were to be investigated, these three characteristics would probably still apply. We can therefore fit a **theoretical distribution** to the waiting times in which these three aspects are included. Table 10.2 reflects the empirical and theoretical distributions of waiting times. In the last column of this table $p(x)$ denotes the probability that X will take on the value of x, i.e. $p(x) = P(X = x)$.

Table 10.2 Empirical and theoretical distributions of waiting times of customers for repairs to electrical appliances

Number of weeks x	Empirical distribution f_x/n	Theoretical distribution, $p(x)$
1	0.075	0.083
2	0.150	0.167
3	0.225	0.250
4	0.283	0.250
5	0.175	0.167
6	0.092	0.083
	1.000	1.000

From the above example it is clear that there is an analogy between the relative empirical distribution of the sample and the theoretical distribution that has been fitted to the data. Although the latter does not necessarily reflect the situation **exactly**, it is very useful if an acceptable theoretical distribution can be postulated in a specific situation. This discussion illustrates the important concept of probability models.

A **probability model** is an idealized and simplified representation of the distribution of a stochastic variable.

A probability model expresses the theoretical distribution of a stochastic variable in a mathematical formula. In the case of a discrete stochastic variable the mathematical expression is known as the **probability distribution**, and in the case of a continuous stochastic variable we refer to the **density function**. We can, for example, express the probability distribution for the discrete stochastic variable X = number of weeks' waiting time in Example 10.1 as follows:

$$p(x) = P(X = x) = \begin{cases} \frac{x}{12} & \text{for } x = 1, 2, 3 \\ \frac{7 - x}{12} & \text{for } x = 4, 5, 6 \end{cases}$$

This probability distribution is represented graphically in Figure 10.1.

A continuous stochastic variable is studied in Example 10.2.

EXAMPLE 10.2

Consider the stochastic variable Y, the price per square metre of residences in Johannesburg. Table 3.7 gives the relative frequency distribution of a random sample of size 100 from this population, and the histogram of the frequency distribution is represented in Figure 3.2. Table 10.3 reflects the frequency distribution and relative frequency distribution of the prices per square metre for the total population of 843 residences if ten intervals of length R200 each are used. Figure 10.2 is a graph of the histogram and frequency polygon of this distribution. It should be emphasized that this frequency distribution and graphic representation

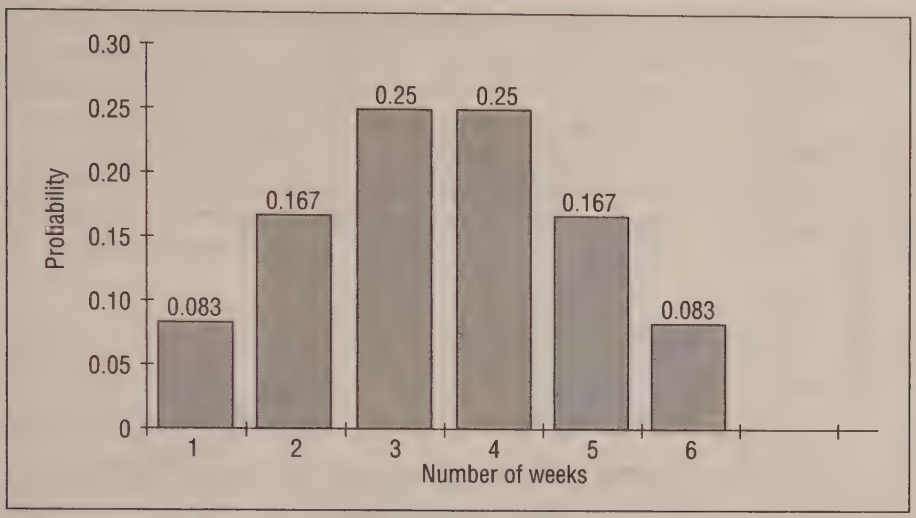


Figure 10.1 Probability distribution of X, the waiting periods of customers for repairs to electrical appliances

are not unique to the distribution of Y, but give only an approximate representation of the distribution, since the choice of the number of classes and the class length also influence the distribution.

Table 10.3 Frequency distribution and relative frequency distribution of the price per square metre of a population of 843 residences in Johannesburg (see Appendix A)

Class interval	Frequency	Relative frequency
[100, 300)	17	0.0202
[300, 500)	72	0.0854
[500, 700)	156	0.1851
[700, 900)	174	0.2064
[900, 1 100)	173	0.2052
[1 100, 1 300)	121	0.1435
[1 300, 1 500)	82	0.0973
[1 500, 1 700)	35	0.0415
[1 700, 1 900)	9	0.0107
[1 900, 2 100)	4	0.0047
	843	1.0000

Figure 10.2 shows clearly that, despite the fact that a small population is involved, the histogram and frequency polygon have a smoother shape than is the case for the random sample of 100 values from this population (see Figure 3.2). As the population size increases and the class lengths become shorter, the histogram and the frequency polygon will blend until both eventually become a smooth curve. Figure 10.3 shows such a smooth curve for the distribution of

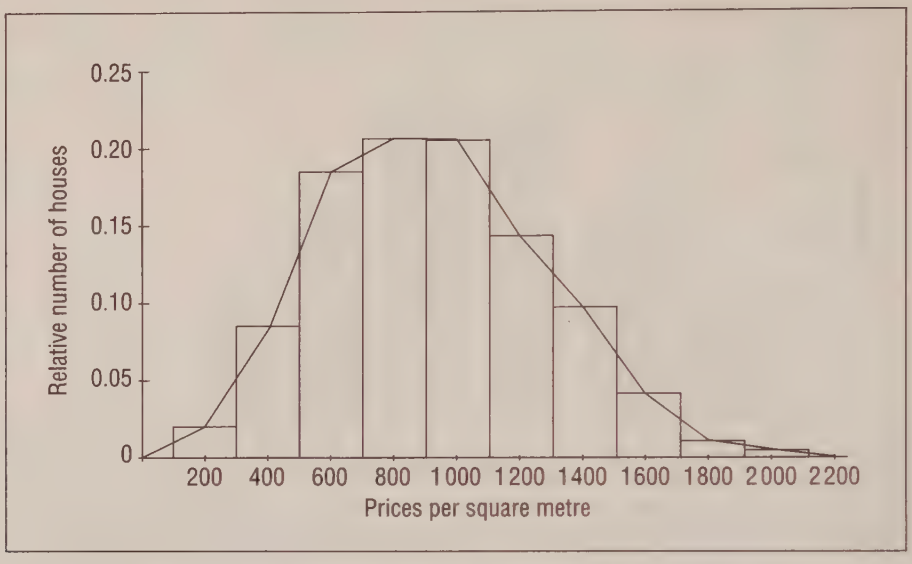


Figure 10.2 Histogram and frequency polygon of the relative frequency distribution of the price per square metre of a population of 843 residences in Johannesburg

house prices per square metre. The scale on the vertical axis was selected so that the total area below the curve is equal to 1. This smooth curve is the density function referred to in the discussion prior to this example and the function $f(y)$ describes the curve mathematically.

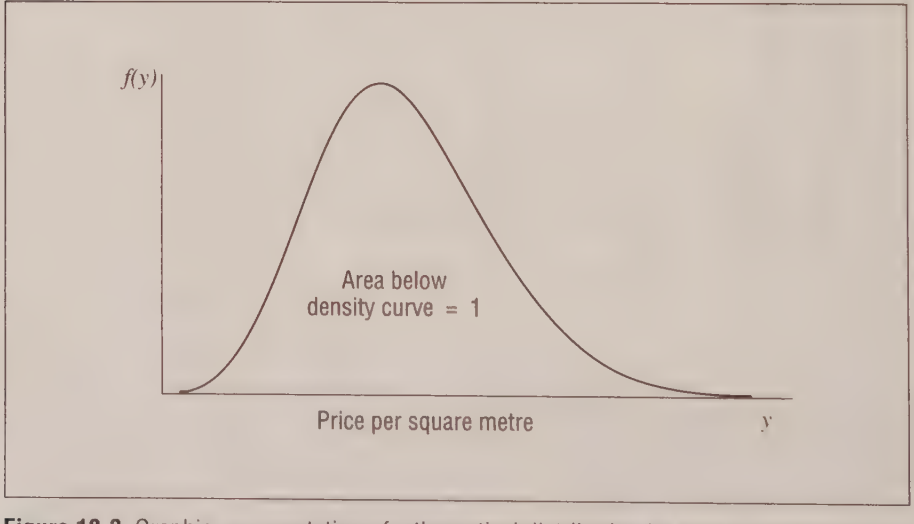


Figure 10.3 Graphic representation of a theoretical distribution by using the density function

Actually, the function $f(y)$ is a probability model and is thus an **idealized** representation of the distribution of the variable Y. The role of the function $f(y)$ is analogous to that of $p(x)$ in the case of a discrete variable. However, it is important to note that $f(y)$ does not represent a probability expression, but is nonetheless used to determine probabilities for the distribution of the stochastic variable Y. The mathematical techniques of area determination are used to determine probabilities for a continuous stochastic variable. This is illustrated in the following example.

EXAMPLE 10.3

A passenger wants to use the suburban bus service but can only remember that there is a bus every twenty minutes at the bus stop where he wants to board. The arrival times are therefore unknown. If he arrives at the bus stop at a given time, he would, for example, like to know what the probability is that he will have to wait

(a) longer than ten minutes;
(b) longer than eight minutes but less than twelve minutes.

We can represent the passenger's waiting time as a continuous variable T with domain $R_T = \{t \mid 0 \leq t \leq 20\}$. We can also assume that the time of his arrival at the bus stop is a randomly selected point between 0 and 20. The density function $f(t)$ of T is represented in Figure 10.4.

Note that $f(t)$ is a horizontal line between $t = 0$ and $t = 20$, with

$$f(t) = \frac{1}{20}, \quad 0 \leq t \leq 20.$$

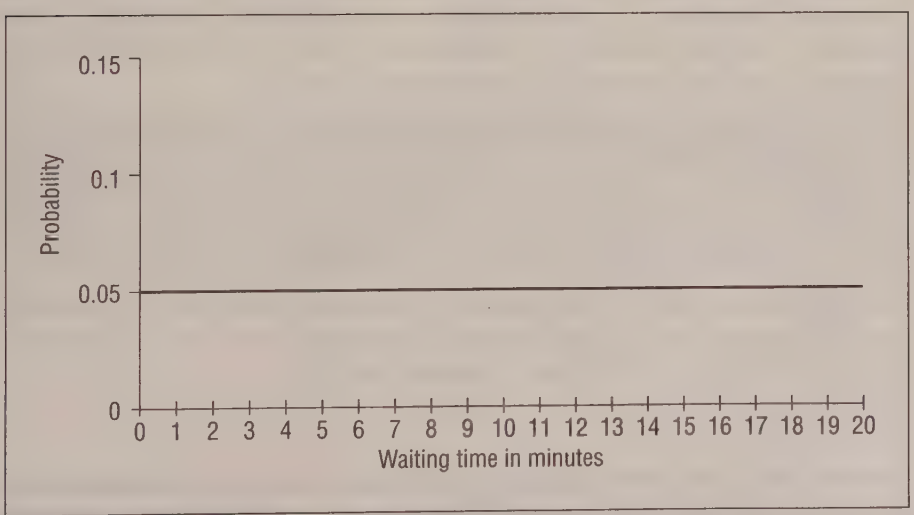


Figure 10.4 Density function of waiting time T of a passenger at a bus stop

The answers to the above questions can be given by:

(a) $P(T > 10)$
= area below the straight line between 10 and 20
= $10 \times \frac{1}{20}$
= 0.5

(b) $P(8 < T < 12)$
= area below the straight line between 8 and 12
= $4 \times \frac{1}{20}$
= 0.2

The above example illustrates the concept of the probability model in the case of a continuous stochastic variable. The probability model in Example 10.3 is known as the **uniform model**. It is also sometimes called the distribution of 'limited information', since the horizontal density function allocates an equal probability to intervals of equal length on the x axis.

In the case of continuous stochastic variables, the probabilities are calculated by determining the areas of appropriate sections below the density function, as illustrated in the above example. In many cases the density function is such that the areas cannot be determined by simple geometrical methods and the more advanced technique of integral calculus has to be applied. However, this technique is beyond the scope of this book.

Further important discrete and continuous probability models and their characteristics are discussed in Sections 3 to 9.

2.2 Cumulative distributions

As has already been shown, the distributions of discrete and continuous stochastic variables are represented by analogous functions $p(x)$ and $f(x)$ respectively. However, it is also possible to represent the distribution of any stochastic variable (discrete or continuous) by means of a probability expression known as the cumulative distribution function, F.

The **cumulative distribution function** of a stochastic variable X is defined by

$F(b) = P(X \leq b)$

for every real value of b.

An important feature of the cumulative distribution function F is that any probability expression of the stochastic variable concerned can be determined by means of F. Thus, for example, it follows that:

$P(X > a) = 1 - F(a)$
$P(a < X \leq b) = F(b) - F(a)$

However, the cumulative distribution functions of discrete and continuous stochastic variables do differ in some important aspects, as the following examples and discussion will illustrate.

EXAMPLE 10.4

Values of the cumulative distribution function of
X = number of weeks' waiting time
of customers who hand in electrical appliances for repairs are calculated as follows (see Example 10.1 and specifically Table 10.2):

$$\begin{aligned} F(1) &= P(X \leq 1) \\ &= P(X = 1) \\ &= 0.083 \end{aligned}$$

$$\begin{aligned} F(2) &= P(X \leq 2) \\ &= P(X = 1) + P(X = 2) \\ &= 0.083 + 0.167 \\ &= 0.250 \end{aligned}$$

$$\begin{aligned} F(3) &= P(X \leq 3) \\ &= P(X = 1) + P(X = 2) + P(X = 3) \\ &= 0.083 + 0.167 + 0.250 \\ &= 0.500, \text{ etc.} \end{aligned}$$

We also note that

$$\begin{aligned} F(1.5) &= P(X \leq 1.5) \\ &= P(X \leq 1), \text{ because } X \text{ can only assume the values } 1, 2, \ldots, 6 \\ &= F(1) \\ &= 0.083 \end{aligned}$$

Similarly, for example,

$$\begin{aligned} F(3.75) &= F(3) \\ &= 0.5 \end{aligned}$$

On the left side of Table 10.4, $F(b)$ is given for $b = 1, 2, \ldots, 6$. On the right side of the table $F(b)$ is given for every real value of b. Note that $F(b)$ remains constant at 0 for all b values in the interval $(-\infty, 1)$, at 0.083 for all b values in the interval $[1, 2)$, and so forth. Figure 10.5 is a graphic representation of $F(b)$ for all real values of b between $-\infty$ and $+\infty$.

Table 10.4 Cumulative distribution function of X, the number of weeks' waiting time of customers for repairs to electrical appliances

b	$F(b)$	b element of	$F(b)$
0	0	$(-\infty, 1)$	0
1	0.083	$[1, 2)$	0.083
2	0.250	$[2, 3)$	0.250
3	0.500	$[3, 4)$	0.500
4	0.750	$[4, 5)$	0.750
5	0.917	$[5, 6)$	0.917
6	1.000	$[6, \infty)$	1.000

The cumulative distribution function of any discrete stochastic variable is a step function with step heights equal to $p(x)$, the probabilities of the various values. It is also important to note that $F(b)$ and probability expressions such as

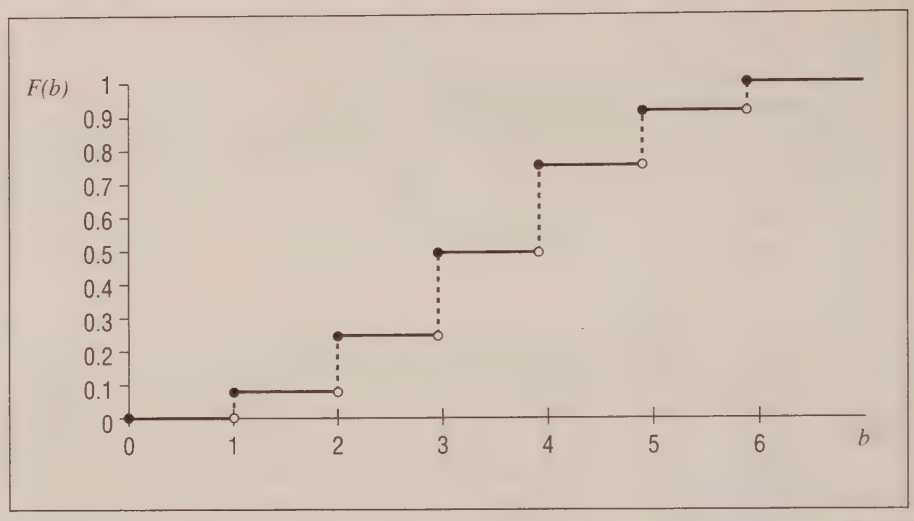


Figure 10.5 Cumulative distribution function, $F(b)$, for X = waiting time in weeks of customers for repairs to electrical appliances

$P(a < X \leq b)$ for any values of a and b, $-\infty < a < b < \infty$, can be determined directly from the graphic representation of $F(b)$. For example, it follows directly from Figure 10.5 that for X = waiting time in weeks

$$\begin{aligned} F(4.5) &= 0.750 \\ P(1.5 < X \leq 5.5) &= P(X \leq 5.5) - P(X \leq 1.5) \\ &= F(5.5) - F(1.5) \\ &= 0.917 - 0.083 \\ &= 0.834 \end{aligned}$$

We note in Example 10.3 above that in the case of continuous stochastic variables, probabilities can be calculated by means of techniques of area determination. As the cumulative distribution function is a probability expression, it can also, in the case of continuous stochastic variables, be determined by means of techniques of area calculation. In some cases simple geometrical methods are adequate, as illustrated in the next example.

EXAMPLE 10.5

Consider the case of the passenger who wants to use the suburban bus service discussed in Example 10.3, and let

T = waiting time until the next bus arrives.

Now, for $0 < t < 20$

$$\begin{aligned} F(t) &= P(T \leq t) \\ &= \text{area below density function between 0 and } t \\ &= t \times \tfrac{1}{20} \end{aligned}$$

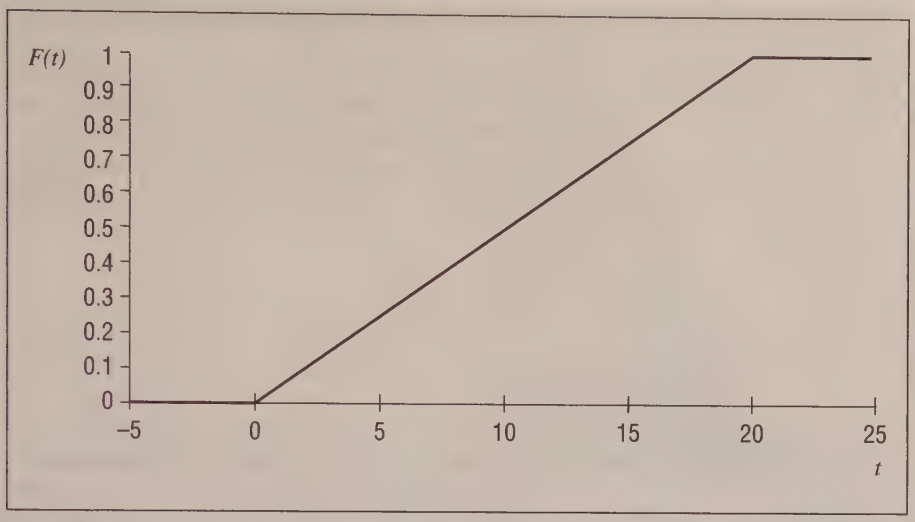


Figure 10.6 Cumulative distribution function $F(t)$ for T = waiting time at bus stop

A general expression for the cumulative distribution function is therefore

$$F(t) = \begin{cases} 0, & t \leq 0 \\ \frac{t}{20}, & 0 < t < 20 \\ 1, & t \geq 20 \end{cases}$$

Graphically, this cumulative distribution function can be represented as indicated in Figure 10.6.

The cumulative distribution function of the continuous stochastic variable Y is defined in Section 2.1 and discussed in Example 10.2. This function can be determined by means of the mathematical technique of integral calculus only if the functional form of the density function $f(y)$ is known or can be approximated. Since this technique is beyond the scope of this book, it suffices to note that, on the basis of Figure 10.7:

> $F(b) = P(Y \leq b)$ is the area to the left of the point b below the density function (dotted section in Figure 10.7).

As seen from Figures 10.2 and 10.3, the density function is actually the 'histogram' of the population, but so scaled that the total area below the curve is 1. With this analogy between the histogram of the sample and the density function in mind, it is also clear that

> $P(b \leq T \leq c) = F(c) - F(b)$, the area below the density function between $y = b$ and $y = c$ (hatched area in Figure 10.7).

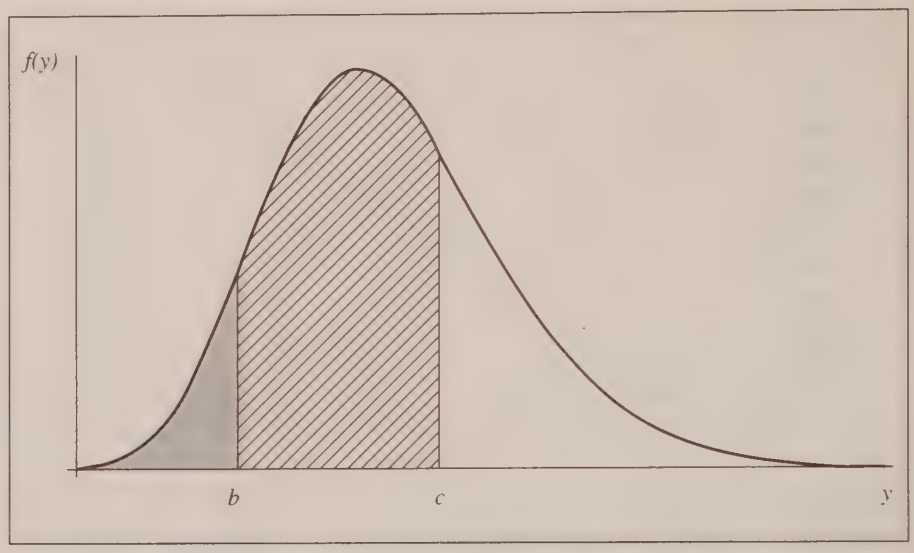


Figure 10.7 Interpretation of probabilities in terms of areas below the density function

Note that the cumulative distribution function of any continuous stochastic variable is a continuous, non-descending function, and in contrast to the cumulative distribution function of a discrete stochastic variable, it shows no 'jumps' (compare, for example, Figures 10.5 and 10.6). It is also important to note that $F(b)$ and mathematical expressions such as $P(a < X \leq b)$ for any values of a and b, $-\infty < a < b < \infty$ can be determined directly from the graphic representation of $F(b)$. For example, it follows directly from Figure 10.6 that:

$$\begin{aligned} F(12) &= 0.6 \\ \text{and } P(8 < T \leq 18) &= F(18) - F(8) \\ &= 0.9 - 0.4 \\ &= 0.5 \end{aligned}$$

Finally, we note from the continuous nature of the density function that

$$\begin{aligned} P(Y \leq b) &= P(Y < b) \\ &= \text{area below density function to the left of } b \end{aligned}$$

because the area is zero at a single point. The probability of a continuous stochastic variable assuming any specific value is therefore theoretically nil.

There is a considerable variety of probability models and each probability model has a unique cumulative distribution function $F(b)$. In practical situations certain probability models are more applicable than others. For this reason these important probability models have tables from which the values of $F(b)$ can be read for different values of b. This aspect is discussed in greater detail in the next sections of this chapter.

2.3 The mean and standard deviation of a stochastic variable

Just as frequency or empirical distributions are described by measures of locality and spread, theoretical distributions of stochastic variables can be described by similar measures of locality and spread.

According to Chapter 4, Section 3, the arithmetic mean of a grouped data set can be calculated from

$$\bar{x} = \frac{\sum fx}{n}$$
$$= \sum x\left(\frac{f}{n}\right)$$

The arithmetic mean $\bar{x}$ can therefore also be calculated from the sum of products of the x values and their respective relative frequencies. From the analogy between empirical and theoretical distributions it follows that the mean of a theoretical distribution, or of the associated stochastic variable, can be calculated from the sum of products of values and their respective probabilities. In order to distinguish between the arithmetic mean, $\bar{x}$, of an observed set of values and the mean of a theoretical distribution, the latter mean is designated by the Greek symbol μ.

The mean or expected value of a discrete stochastic variable X is defined by

$$\mu = \sum xp(x)$$

and is known as the **population mean**.

(The reader is also referred to Chapter 4, Section 2, where the mean μ for a population of N elements is defined as follows: $\mu = \frac{1}{N}\sum x$. For purposes of this book these two definitions are regarded as equivalent.)

EXAMPLE 10.6

Suppose we want to calculate the mean number of weeks' waiting time of a customer for repairs to electrical appliances (see Section 2.1 and Example 10.1). The theoretical distribution of X = waiting time in weeks is given by

x	1	2	3	4	5	6
$p(x)$	0.083	0.167	0.250	0.250	0.167	0.083

so that the mean or expected value μ is given by

$$\begin{aligned}\mu &= \Sigma x p(x) \\ &= 1 \times 0.083 + 2 \times 0.167 + \ldots + 6 \times 0.083 \\ &= 3.5\end{aligned}$$

The arithmetic mean of the sample of 120 waiting times (see Table 10.1) is

$$\begin{aligned}\bar{x} &= \Sigma x\left(\frac{f}{120}\right) \\ &= 1 \times 0.075 + 2 \times 0.150 + \ldots + 6 \times 0.092 \\ &\approx 3.6\end{aligned}$$

If a second random sample of 120 values is taken from the population of x values, the $\bar{x}$ value will probably differ from 3.6. On the other hand, the population mean number of weeks' waiting time for customers is a constant, namely $\mu = 3.5$.

According to Chapter 5, Section 3, the standard deviation of a grouped data set is given by

$$\begin{aligned}s &= \sqrt{\frac{\sum f(x - \bar{x})^2}{n - 1}} \\ &= \sqrt{\sum (x - \bar{x})^2 \frac{f}{n - 1}}\end{aligned}$$

In the discrete case, the standard deviation of a theoretical distribution or of the associated stochastic variable can now be defined as follows:

The standard deviation of a discrete stochastic variable X is defined by

$$\sigma = \sqrt{\sum (x - \mu)^2 p(x)}$$

and is also known as the **population standard deviation**.

We note that the relative frequency $\frac{f}{n}$ in the formula of $\bar{x}$ is replaced by $p(x)$ in the definition of μ. This corresponds with the frequency limit definition of probability. However, in the formulas for s and σ, $\frac{f}{n-1}$ and $p(x)$ have the same role. As explained in Chapter 5, Section 2, we obtain a more accurate estimate of σ from s with $\frac{f}{n-1}$ in the definition than would be the case had we used $\frac{f}{n}$ in the definition of s. However, it is clear that $\frac{f}{n}$ and $\frac{f}{n-1}$ do not differ much for large values of n.

330 Analogous to the calculation formula of s, which is expressed as:

$$s = \sqrt{\sum x^2 \frac{f}{n-1} - \frac{n}{n-1}\bar{x}^2}$$

the calculation formula for σ is:

$$\sigma = \sqrt{\sum x^2 p(x) - \mu^2}$$

EXAMPLE 10.7

The population and sample standard deviations of X = number of weeks' waiting time for repairs to electrical appliances is:

$$\begin{aligned}\sigma &= \sqrt{\sum x^2 p(x) - \mu^2} \\ &= \sqrt{\{1^2 \times 0.083 + 2^2 \times 0.167 + \ldots + 6^2 \times 0.083\} - 3.5^2} \\ &= \sqrt{14.164 - 12.25} \\ &\approx 1.38\end{aligned}$$

$$\begin{aligned}s &= \sqrt{\sum x^2 \frac{f}{n-1} - \frac{n}{n-1}\bar{x}^2} \\ &= \sqrt{\left\{1^2 \times \frac{9}{119} + 2^2 \times \frac{18}{119} + \ldots + 6^2 \times \frac{11}{119}\right\} - \frac{120}{119} \times 3.6^2} \\ &= \sqrt{15.03361 - 13.06891} \\ &\approx 1.40\end{aligned}$$

Notes

(a) The mean and standard deviation of a continuous stochastic variable can be calculated only if the density function $f(x)$ is known. This is achieved by means of the mathematical technique of integration, namely

$$\mu = \int_{-\infty}^{+\infty} x f(x) dx$$

$$\sigma = \sqrt{\int_{-\infty}^{+\infty} x^2 f(x) dx - \mu^2}$$

In the case of a stochastic variable with a symmetrical distribution, the mean is always the point of symmetry. Thus for T = waiting time at bus stop (Example 10.3)

μ = expected waiting time
= 10 minutes

(b) It is standard practice to indicate descriptive measures for samples with Latin symbols ($\bar{x}$, s, and so forth) and the corresponding descriptive measures for theoretical distributions with Greek symbols (μ, σ, and so forth).

(c) Measures of symmetry and kurtosis for theoretical distributions can also be defined on the analogy of the measures for empirical distributions.

3 The binomial distribution

Data sets observed in practice can often be seen as realizations of some well-known stochastic variable X. The exact probability distribution of X (discrete case) or the mathematical form of the density function (continuous case) is often not known. However, the circumstances under which the stochastic variable was observed, together with the characteristics of the observed x values, usually identify an appropriate probability model.

In this section and subsequent sections certain important discrete and continuous probability models are discussed.

Certain stochastic experiments consist of the repeated execution of a certain action. Thus, for example, the experiment in Example 9.1 consists of tossing a buckled 1c coin fifty times and recording the number of times that heads come up. Consider the following more practical situation: In a certain manufacturing process each completed article is classified as defective or non-defective. The quality manager who supervises the process knows from experience that if the process is under control, 5% of the articles are defective. In order to monitor the process, he has to draw a random sample of twenty articles from a batch of articles (the repeated action of selecting and testing one article). It is important for the manager to be able to assess the probability of, for example, more than two defective articles being included in a random sample of twenty articles.

The above example is a typical application of the **binomial experiment**. This experiment is characterized by the following features:

(a) The experiment consists of a fixed number, say n, repetitions of an action, called **trials** (also known as **Bernoulli trials**).

(b) The different trials are statistically independent of each other.

(c) The outcome of each trial is **dichotomous**, i.e. the outcome of each trial can be one of only two possibilities. These two possibilities are usually indicated by the terms **success** and **failure**.

(d) In each trial the **probability of success is the same**, say π.

(e) The **number of successes** attained in the n trials is noted.

It is clear that the stochastic variable X = number of successes attained in a binomial experiment with n trials is a discrete variable. This variable has a probability distribution known as the **binomial distribution with parameters** n and π, and it is written briefly as the bi(n, π) distribution.

The probability distribution of X, the number of successes in a binomial experiment, is given by

$$p(x) = \binom{n}{x}\pi^x(1 - \pi)^{(n-x)};\ x = 0,\ 1,\ 2,\ \ldots,\ n.$$

This expression is used at the end of Chapter 9 as an approach to the situation in which a sample of size n is drawn from a population of size N with $n < < N$ (see Example 9.14). This result, however, will apply exactly if **sampling with replacement** occurs, in other words when each article drawn is replaced once it has been inspected and before the next one is drawn. In Example 10.8 such a situation is simulated by generating a sample of random numbers.

It should be noted that the mean and standard deviation of the binomial distribution are given by

$$\mu = n\pi \text{ and } \sigma = \sqrt{n\pi(1 - \pi)}.$$

(see Note (c) below).

EXAMPLE 10.8

A stochastic experiment consists of drawing six random numbers and noting the number of even numbers X, with 0 also being regarded as an even number. It is clear that this experiment is indeed a binomial experiment with $n = 6$ and $\pi = 0.5$.

The experiment is now carried out 500 times and each time the value of X is noted. The relative frequencies of the number of times out of 500 that 0, 1, ..., 6 successes are attained are recorded, as is the probability distribution of X, namely $p(x)$, for $x = 0,\ 1,\ \ldots,\ 6$. A summary is given in Table 10.5.

Table 10.5 Relative frequencies and probability distribution of X, the number of even numbers among six random numbers, obtained from 500 executions of the binomial experiment

Number of even numbers, x	Frequency of x	Relative frequency of x	Probability distribution of X, $p(x)$
0	12	0.024	0.015625
1	28	0.056	0.093750
2	123	0.246	0.234375
3	156	0.312	0.312500
4	127	0.254	0.234375
5	42	0.084	0.093750
6	12	0.024	0.015625
	500	1.000	1.000000

From the definition of the bi(6, 0.5) distribution it follows that

$$p(x) = \binom{6}{x}(0.5)^x(0.5)^{6-x}$$
$$= \binom{6}{x}(0.5)^6, \quad x = 0, 1, 2, \ldots, 6$$

Thus, for example,

$$p(0) = \binom{6}{0}(0.5)^6$$
$$= 0.015625$$
$$p(1) = \binom{6}{1}(0.5)^6$$
$$= 0.093750$$

Generally speaking, there is a fairly pronounced similarity between the empirical distribution (third column) and the theoretical distribution (last column) as seen in Table 10.5. The probability distribution of the bi(6, 0.5) distribution is represented graphically in Figure 10.8 by means of a bar chart. The mean and standard deviation of X are given by

$$\mu = n\pi \quad \text{and} \quad \sigma = \sqrt{n\pi(1-\pi)}$$
$$= 6(0.5) \qquad = \sqrt{6(0.5)(0.5)}$$
$$= 3 \qquad = 1.22$$

Notes

(a) As clearly revealed by the last column of Table 10.5 and also by Figure 10.8, the bi(6, 0.5) distribution is symmetrical. This feature applies to any binomial distribution for which $\pi = 0.5$. For a fixed value of n, however,

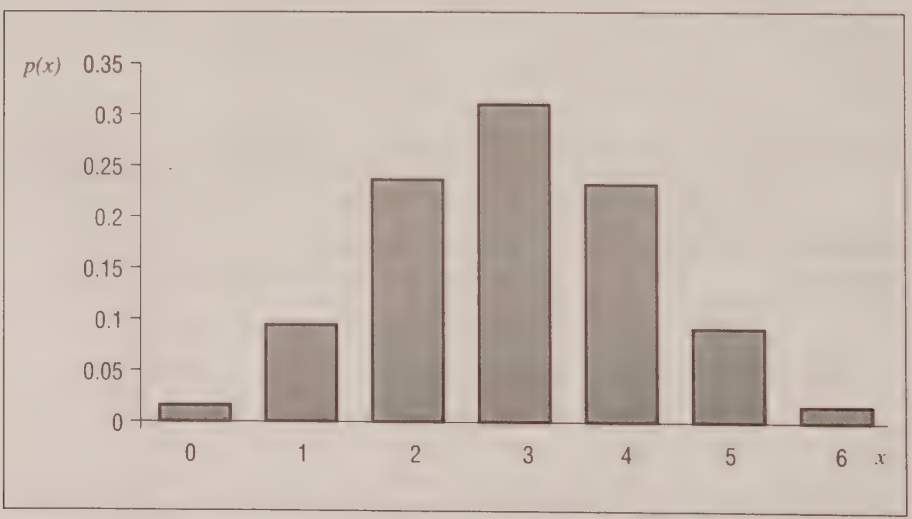


Figure 10.8 Bar chart of the probability distribution of the bi(6, 0.5) distribution

the binomial distribution becomes increasingly skewed as the difference between π and 0.5 increases. On the other hand, the greater the value of n, the more symmetrical the binomial distribution, regardless of the difference between π and 0.5.

(b) The derivation of the probability expression for the binomial distribution is briefly as follows:

Obtaining x successes in n trials can manifest in different ways, for example *SS ... SMM ... M* or *SMSM ...* et cetera, in which x S symbols (denoting 'success') and $(n - x)$ M symbols (denoting 'failure') are found in the row. Since the trials are mutually independent, it follows that

$$\begin{aligned} P(SS \ldots SMM \ldots M) &= \pi\pi \ldots \pi(1-\pi)(1-\pi) \ldots (1-\pi) \\ &\quad (\pi \text{ } x \text{ times and } 1-\pi \text{ } (n-x) \text{ times}) \\ &= \pi^x(1-\pi)^{n-x} \end{aligned}$$

Similarly,

$$\begin{aligned} P(SMSM \ldots) &= \pi(1-\pi)\,\pi(1-\pi) \ldots \\ &\quad (\pi \text{ } x \text{ times and } 1-\pi \text{ } (n-x) \text{ times}) \\ &= \pi^x(1-\pi)^{n-x} \end{aligned}$$

We thus find that all combinations of x *S* symbols and $(n - x)$ *M* symbols have the same probability, namely $\pi^x(1 - \pi)^{n-x}$. As there are $\binom{n}{x}$ different combinations of x *S* symbols and $(n - x)$ *M* symbols,

$$\begin{aligned} P(X = x) &= P(x \text{ } S\text{s and } (n-x) \text{ } M\text{s}) \\ &= P(SS \ldots SMM \ldots M) + P(SMSM \ldots) + \ldots \quad (\tbinom{n}{x} \text{ terms}) \\ &= \pi^x(1-\pi)^{n-x} + \pi^x(1-\pi)^{n-x} + \ldots \quad (\tbinom{n}{x} \text{ terms}) \\ &= \tbinom{n}{x}\pi^x(1-\pi)^{n-x} \end{aligned}$$

(c) The binomial distribution derives its name from the well-known binomial expansion of the nth power of a sum:

$$\begin{aligned} (a+b)^n &= \tbinom{n}{0}a^0b^n + \tbinom{n}{1}a^1b^{n-1} + \tbinom{n}{2}a^2b^{n-2} + \ldots + \tbinom{n}{n}a^nb^0 \\ &= \sum_{x=0}^{n}\tbinom{n}{x}a^xb^{n-x} \end{aligned}$$

The application of this expansion yields two important results:

(i) $$\sum_{x=0}^{n} p(x) = \sum_{x=0}^{n}\tbinom{n}{x}\pi^x(1-\pi)^{n-x} = \left\{\pi + (1-\pi)\right\}^n = 1$$

(ii) $$\begin{aligned} \pi &= \sum_{x=0}^{n} xp(x) \\ &= \sum_{x=1}^{n} xp(x) \qquad \text{because } 0.\,p(0) = 0 \\ &= \sum_{x=1}^{n} x\tbinom{n}{x}\pi^x(1-\pi)^{n-x} \\ &= n\pi\sum_{x=1}^{n}\tbinom{n-1}{x-1}\pi^{x-1}(1-\pi)^{n-1-(x-1)}, \qquad \text{after simplification} \end{aligned}$$

$$= n\pi \sum_{y=0}^{n-1} \binom{n-1}{y} \pi^y (1-\pi)^{n-1-y}, \qquad \text{with } y = x - 1$$

$$= n\pi \{\pi + (1-\pi)\}^{n-1}$$

$$= n\pi$$

(d) In Sections 4 and 7 we see that the binomial distribution for large values of n can, under certain conditions, be approximated by the Poisson and normal distributions.

3.1 Tables of the cumulative binomial distribution

If X has a bi(n, π) distribution, probability expressions such as $P(X \leq b)$ or $P(a \leq X \leq b)$ often have to be determined. This determination involves the calculation of a whole number of consecutive terms of the form $\binom{n}{x} \pi^x (1-\pi)^{n-x}$. If n is fairly large, the task can be cumbersome and we therefore use tables of the cumulative distribution function of the binomial distribution. Appendix B, Table B.6 reflects $F(b)$ for different values of n, π and b. The following example illustrates the use of this table.

EXAMPLE 10.9

It is a known fact that 10% of all light bulbs manufactured by a certain factory are defective. Suppose we want to determine the probability that in a random sample of fifteen bulbs: (i) at most two; (ii) at least four; and (iii) one to five light bulbs are defective.

Let X be the number of defective light bulbs in the sample of fifteen. It is clear that this is a binomial experiment and that X has a bi(15, 0.10) distribution.

(i) $P(X \leq 2) = F(2)$
$= 0.81594$ (Table B.6 with $n = 15$, $\pi = 0.10$ and $b = 2$)

(ii) $P(X \geq 4) = 1 - P(X \leq 3)$
$= 1 - F(3)$
$= 1 - 0.94444$ (Table B.6 with $n = 15$, $\pi = 0.10$ and $b = 3$)
$= 0.05556$

(iii) $P(1 \leq X \leq 5) = F(5) - F(0)$
$= 0.99775 - 0.20589$
$= 0.79186$

4 The Poisson distribution

In some binomial experiments it is neither meaningful nor possible to count all trials. Consider the following two apt examples:

(a) Vehicles move through an intersection of two busy streets. There is a daily potential for many accidents ('trials' of a binomial experiment), but the actual number of accidents that do take place ('successes' in the experiment) is relatively small.

(b) The number of fire alarm calls received daily by a fire station in a large city can be regarded as the successes in a binomial experiment. However, the failures and therefore the total number of trials can be neither identified nor counted.

In binomial experiments of this nature the **Poisson probability model** provides a simple approach to the probability distribution of X, the number of successes. This type of experiment is characterized by the following properties:

(a) There is ample opportunity for a success to occur, but relatively few successes are realized. Successes can therefore be regarded as 'rare events'.

(b) The probability of a success being realized in a measurement interval of a certain length is the same for all measurement intervals of the same length, i.e. the incidence of successes is random over the measurement interval.

(c) The incidences of successes are independent of one another.

Measurement interval in (c) above can refer to time intervals but also to, for instance, distances and areas. Thus, for example, the determination of X, the number of defects per square metre in clothing material, can also be regarded as the execution of a Poisson experiment. The **Poisson distribution** has only one parameter, namely λ, the mean number of successes in a measurement interval of unit length. The abbreviated form is the **po(λ) distribution**.

The probability of obtaining X successes in a measurement interval of size t for the po(λ) distribution is given by

$$p_t(x) = P(X = x) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}, \quad x = 0, 1, 2, \ldots$$

If X has a Poisson distribution with the probability function $p_t(x)$, the mean and standard deviation of X are given by

$$\mu = \lambda t \quad \text{and} \quad \sigma = \sqrt{\lambda t}$$

EXAMPLE 10.10

A fire station receives on average three fire alarm calls per normal weekday (holidays excluded). The station has five fire engines, each with its own fire-fighting team. With a view to scheduling service times and maintenance of equipment, it is important to find solutions for the following problems:

(a) What is the probability of there being exactly two fire alarm calls on a specific day?

(b) What is the probability of there being more than five fire alarm calls within a period of three hours?

(c) How many alarms are received, on average, per six-hour period and per working week (Monday to Friday)?

Let X = number of fire alarms per day.

In accordance with the above discussion, X is po(3) distributed for a measuring unit of one day. The three questions can therefore be answered as follows:

(a) $P(X = 2) = e^{-3}\frac{3^2}{2!}$

$= 0.22404$

(b) A three-hour period is equivalent to a measurement interval of

$t = \frac{1}{8}$ day so that $\lambda t = 3 \times \frac{1}{8} = 0.375$

Therefore $P(X > 5) = 1 - P(X \leq 5)$

$$= 1 - \sum_{x=0}^{5} e^{-0.375}\frac{(0.375)^x}{x!}$$

$$= 1 - e^{-0.375}\left(1 + 0.375 + \frac{(0.375)^2}{2!} + \ldots + \frac{(0.375)^5}{5!}\right)$$

$= 1 - 0.999997$

$= 0.000003$

(c) A six-hour period is equivalent to a measurement interval of

$t = \frac{1}{4}$ day so that $\lambda t = 3 \times \frac{1}{4} = 0.75$

Therefore the mean number of fire alarms per six-hour period is

$\mu = \lambda t = 0.75.$

A week period is equivalent to a measurement interval of

$t = 5$ days and $\lambda t = 3 \times 5 = 15.$

Therefore the mean number of fire alarms per working week is

$\mu = \lambda t = 15.$

4.1 The Poisson approximation of the binomial distribution

In cases where the trials in a binomial experiment can be counted, the Poisson distribution offers a simple approximation of the probability distribution of X, the number of successes, provided that π, the probability of a success, is small and n, the number of trials, is large. Formally, this result can be expressed as follows:

$$\binom{n}{x}\pi^x(1-\pi)^{n-x} \approx e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots, n$$

provided π is small and n is large, and with $\lambda = n\pi$

4.2 Tables of the cumulative Poisson distribution

From Example 10.10 it is clear that, as in the case of the binomial distribution, the sum of consecutive Poisson probabilities often has to be calculated. In Appendix B, Table B.7, $P(X \leq b)$ is tabulated for different values of λ and b, with X distributed according to the po(λ) distribution. The use of the table and the Poisson approximation of the binomial distribution is illustrated by the following example.

EXAMPLE 10.11

A publishing company has printed 10 000 copies of a certain textbook. In the first year of its distribution, 1 400 copies are sold, of which seven are returned because of defects. Determine for a new order of 300 copies:

(a) The probability of no more than three copies being returned

(b) The probability of more than one but fewer than eight copies being returned.

On the basis of the first-year distribution figure and with the aid of the frequency limit definition, the probability of a defect copy can be set at $\pi = \frac{7}{1\,400} = 0.005$. For a future order of 300 copies the binomial experiment is thus applicable, with $n = 300$ and $\pi = 0.005$. From the Poisson approximation of the binomial distribution it follows that:

$$\begin{aligned}\lambda &= n \times \pi \\ &= 300 \times 0.005 \\ &= 1.5\end{aligned}$$

(a) $P(X \leq 3) = F(3)$
$= 0.93436$, by using Table B.7.

(b) $P(1 < X < 8) = P(X \leq 7) - P(X \leq 1)$
$= F(7) - F(1)$
$= 0.99983 - 0.55783$
$= 0.442$

Note

(a) Strictly speaking, the number of defective books in Example 10.11 is distributed according to the **hypergeometric distribution** and the properties of the binomial experiment are only approximately applicable – see Section 5 below.

(b) For large values of λ the po(λ) distribution can be approximated by the normal distribution – see Section 8.

5 The hypergeometric distribution

As in the case of the binomial and Poisson experiments, the stochastic variable of interest in the **hypergeometric experiment** is the number of successes that are observed. The experiment is illustrated at the end of Chapter 9 in Example 9.13. Characteristics of the hypergeometric experiment are:

(a) The experiment entails drawing n, a fixed number of objects, from a finite population of N objects.

(b) If a particular characteristic is of interest, the N objects can be divided into two groups, namely, N_1 elements with the characteristic and N_2 elements without the characteristic, where $N = N_1 + N_2$. When drawing one element, finding an object with the characteristic concerned is a 'success', while an element without the characteristic is a 'failure'.

(c) The results of the n different draws are dependent on one another in the sense that the result of any draw is influenced by the results of all previous draws.

(d) The probability of a success is not the same from one trial to the next.

(e) X, the number of successes obtained in the n drawings is noted.

The stochastic variable, X, associated with such a hypergeometric experiment has a discrete probability distribution, known as the **hypergeometric distribution**, with parameters n, N_1 and N. This distribution is indicated as the hi(n, N_1, N) distribution.

The probability distribution of X, the number of successes obtained in a hypergeometric experiment, is given by

$$p(x) = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{n}}; \quad x = 0, 1, 2, \ldots, n$$

The mean and standard deviation of the hypergeometric distribution in terms of the notation given above are

$$\mu = n\pi \quad \text{and} \quad \sigma = \sqrt{n\pi(1-\pi)\left(1 - \frac{n-1}{N-1}\right)}$$

with $\pi = \frac{N_1}{N}$

EXAMPLE 10.12

Refer to Example 9.13, where X = number of trucks with low-grade sunflower seed has an hi(5, 12, 40) distribution. For the two questions asked there it follows that

(a) P (exactly two trucks of the lower grade)
$= P(X = 2)$

$$= \frac{\binom{12}{2}\binom{28}{3}}{\binom{40}{5}}$$

$$\approx 0.33$$

(b) $P(\text{at most two trucks of the lower grade})$

$$= P(X \leq 2)$$

$$= \frac{\binom{12}{0}\binom{28}{5}}{\binom{40}{5}} + \frac{\binom{12}{1}\binom{28}{4}}{\binom{40}{5}} + \frac{\binom{12}{2}\binom{28}{3}}{\binom{40}{5}}$$

$$\approx 0.85$$

Furthermore, the mean and standard deviation of X are:

$$\mu = n\pi \qquad \sigma = \sqrt{n\pi(1-\pi)\left(1 - \frac{n-1}{N-1}\right)}$$

$$= 5 \times \frac{12}{40} \qquad = \sqrt{5 \times \frac{12}{40} \times \left(1 - \frac{12}{40}\right) \times \left(1 - \frac{4}{39}\right)}$$

$$= 1.5 \qquad \approx 0.97$$

Notes

(a) In the definition of the hypergeometric distribution the domain is given as $R_X = \{0, 1, 2, \ldots, n\}$. However, strictly speaking,

$$R_X = \{x | x \text{ an integer}, \max(0, n - N_2) \leq x \leq \min(n, N_1)\}.$$

(b) As mentioned at the end of Chapter 9, the hypergeometric probability expression can be approximated by using the binomial probability expression, provided n is much smaller than N:

$$\frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{n}} = \frac{\binom{N\pi}{x}\binom{N(1-\pi)}{n-x}}{\binom{N}{n}} \quad \text{with} \quad \pi = \frac{N_1}{N}$$

$$\approx \binom{n}{x}\pi^x(1-\pi)^{n-x}, \quad \text{provided} \quad n \leq (0.05)N$$

6 The exponential distribution

Probability models for discrete stochastic variables are discussed in the previous three sections. The models dealt with are important because they provide good approximations of distributions that are often found in practice. In the next four sections we consider special continuous probability models. Knowledge of these theoretical models is an essential prerequisite for a good understanding of statistical inference, which is discussed in the next few chapters.

The simplest continuous distribution is the so-called **uniform distribution**, which was discussed as an example in Section 2 and for which the mathematical function of the density function is given by

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise} \end{cases}$$

For the uniform distribution it can be proved that

$$\mu = \frac{a+b}{2} \quad \text{and} \quad \sigma^2 = \frac{(b-a)^2}{12}$$

The Poisson experiment (see Section 4) is often studied by observing the size of measurement periods between consecutive realizations of successes. Thus, for example, the distribution of the waiting time, T, between consecutive fire alarm calls at a fire station, or the distribution of the distances, A, between two consecutive potholes on a tarred road between two towns, can be observed. The 'waiting time' stochastic variable and the 'distance' stochastic variable both have a continuous distribution, known as the **exponential distribution**.

The density function of the exponential distribution is given by

$$f(x) = \tfrac{1}{\theta} e^{\frac{-x}{\theta}}, \quad x > 0 \quad \text{and} \quad \theta > 0$$

The cumulative distribution function of the exponential distribution is given by

$$F(x) = 1 - e^{\frac{-x}{\theta}}, \quad x > 0 \quad \text{and} \quad \theta > 0$$

A stochastic variable with the above given density function and cumulative distribution function has an **exp(θ) distribution**. The mean and standard deviation of the exp(θ) distribution are given by $\mu = \theta$ and $\sigma = \theta$.

From this we note that for the exp(θ) distribution $\sigma^2 = \mu^2 = \theta^2$.

The following result gives the relationship between the Poisson experiment and the exponential distribution:

If X = number of successes in a measurement interval of unit length has a po(λ) distribution, then T = length of measurement interval between consecutive successes will have an **exp($\frac{1}{\lambda}$) distribution**.

EXAMPLE 10.13

In Example 10.10 a fire station receives an average of three fire alarm calls per normal working day, so we can say that X = number of fire alarms per day has a po(3) distribution. From the relation between the Poisson and exponential distributions shown above we can say that the time lapse between two consecutive fire alarm calls has an exp($\frac{1}{3}$) distribution, that is, the mean waiting time between consecutive fire alarm calls is $\theta = \frac{1}{3}$ day or eight hours.

The calculation of probability expressions for the exponential distribution using the cumulative distribution function is as follows:

$$\begin{aligned} P(a \leq X \leq b) &= F(b) - F(a) \quad \text{(see Section 2.2)} \\ &= \left(1 - e^{\frac{-b}{\theta}}\right) - \left(1 - e^{\frac{-a}{\theta}}\right) \\ &= e^{\frac{-a}{\theta}} - e^{\frac{-b}{\theta}} \end{aligned}$$

The use of this formula is illustrated in the next example.

EXAMPLE 10.14

An investigation of the waiting times between the arrival of consecutive clients who require a special type of service is important to financial institutions. From a survey conducted at different times and on different days of a particular month, it transpired that, on average, 28 clients per seven-hour working day require this type of service. Determine the probability of the waiting time being

(a) shorter than fifteen minutes;
(b) between thirty minutes and 1 hour; and
(c) longer than 1 hour.

Since the number of clients per seven-hour day is 28, X = number of clients per hour has a po(4) distribution. Consequently T = waiting period between two consecutive clients has an exp($\frac{1}{4}$) distribution, with $f(x) = \frac{1}{0.25} e^{\frac{-x}{0.25}}$. Therefore:

(a) $$\begin{aligned} P(T < 15 \text{ minutes}) &= P(T < 0.25 \text{ hour}) \\ &= F(0.25) \\ &= 1 - e^{\frac{-0.25}{0.25}} \\ &\approx 0.632 \end{aligned}$$

(b) $$\begin{aligned} P(30 \text{ minutes} < T < 1 \text{ hour}) &= P(0.5 < T < 1.0) \\ &= e^{\frac{-0.5}{0.25}} - e^{\frac{-1.0}{0.25}} \\ &= 0.13534 - 0.018316 \\ &\approx 0.117 \end{aligned}$$

(c) $P(T > 1 \text{ hour}) = 1 - F(1.0)$

$$= 1 - \left(1 - e^{\frac{-1.0}{0.25}}\right)$$

$$\approx 0.018$$

Notes

(a) The exponential model is also used for other stochastic phenomena, for instance the income distribution of populations of households.

(b) One of the most important properties of the exponential distribution is the so-called property of 'no memory'. It can namely be proved that

$$P(T > t + s \mid T > t) = P(T > s) \quad \text{where } t \text{ and } s > 0.$$

Applied to Example 10.14, we can, for instance, say that

$$P(T > 1.25 \text{ hours} \mid T > 1 \text{ hour}) = P(T > 0.25 \text{ hours}),$$

that is, given that the waiting time since the arrival of the previous client is already longer than an hour, the probability that the client will be waiting for longer than 75 minutes is the same as the probability that a client will be waiting for longer than fifteen minutes after the arrival of the previous client.

7 The normal distribution

The **normal distribution** is the most important continuous distribution and was studied by the French mathematician A. de Moivre as far back as 1733. Later the famous mathematicians Laplace and Gauss independently proved that the normal distribution is a suitable probability model for measurement errors in the physical sciences. For this reason the normal distribution is also known as the **Gauss distribution**. Examples of the application of the normal distribution are the following:

(a) Preserved vegetables and fruit are tinned automatically. The filling machine is set so that the mean content of all 500 gram tins is actually 500 gm. However, owing to the inherent variation in the process, it is impossible to ensure that the content of each tin is exactly 500 grams.The continuous stochastic variable X = content of a randomly selected tin has a normal distribution with mean $\mu = 500$ grams.

(b) The final marks scored by matriculants of a certain education department in any subject (for example Mathematics) in a certain year has a normal distribution with, for instance, a mean $\mu = 55\%$.

(c) The fuel consumption in litres per 100 km for a car of a certain make and model has a normal distribution with mean $\mu = 8.5$ litres per 100 km.

The density function of the normal distribution with mean μ and standard deviation σ is represented by the function

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < +\infty$$

From this definition it is clear that the two parameters μ and σ^2 (the mean and variance) fully determine the distribution. It is customary to refer to this distribution as the **n(μ, σ^2) distribution**. Figure 10.9 is a graphic representation of the density functions of a n(25, 9) and a n(30, 4) distribution.

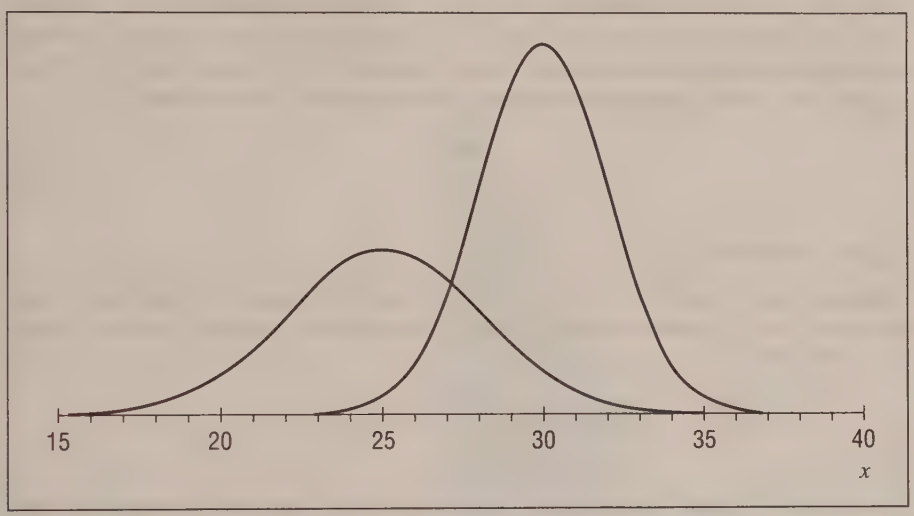


Figure 10.9 The density functions of a n(25, 9) and a n(30, 4) distribution

The following are characteristics of the normal distribution:

(a) The normal curve is bell shaped and symmetrical around the mean μ.

(b) Approximately 68% of the total area below the curve lies between the limits $\mu - \sigma$ and $\mu + \sigma$, approximately 95% of the area lies between $\mu - 2\sigma$ and $\mu + 2\sigma$, and 99.7% of the area lies between $\mu - 3\sigma$ and $\mu + 3\sigma$.

(c) The normal curve approximates the horizontal axis at $+\infty$ and $-\infty$, that is, it is **asymptotic** to the x axis.

In contrast to the situation with uniform and exponential distributions, probabilities for the normal distribution are mainly determined by means of the **standardization** technique and special tables.

If X has the $n(\mu, \sigma^2)$ distribution, the **standard normal variable**

$$Z = \frac{X - \mu}{\sigma}$$

has the n(0, 1) distribution. This n(0, 1) distribution is known as the **standard normal distribution**.

It now follows from characteristic (b) that 68% of the total area below the density function of the n(0, 1) distribution lies between –1 and +1; 95% of the area between –2 and +2; and 99.7% of the area between –3 and +3.

If X is distributed n(25, 9), for example, $Z = \frac{X - 25}{3}$ is distributed n(0, 1), and for Y distributed n(30, 4), $Z = \frac{Y - 30}{2}$ is also distributed n(0, 1). The fact that any normal distribution can be transformed to the standard normal distribution makes it possible to set up one table for determining probability expressions for all normal distributions. Example 10.15 explains this.

EXAMPLE 10.15

Suppose the probability of the n(25, 9) stochastic variable X assuming a value between 22 and 28 needs to be determined. The standard values of these two values are respectively

$$z_1 = \frac{22 - 25}{3} \quad \text{and} \quad z_2 = \frac{28 - 25}{3}$$
$$= -1 \qquad\qquad = 1$$

The required probability in terms of the variable X can now be expressed and calculated as follows in terms of the standardized variable Z:

$$P(22 < X < 28) = P(-1 < Z < +1)$$
$$= 68\%, \text{ as shown above.}$$

However, suppose we need to determine the following probability:

$$P(25 < X < 30) = P\left(\frac{25 - 25}{3} < \frac{X - 25}{3} < \frac{30 - 25}{3}\right)$$
$$= P(0 < Z < 1.67)$$

The latter probability can be determined only by means of tables of the normal distribution.

7.1 Tables of the standard normal distribution

Tables of probabilities of the standard normal distribution are given in different ways in the literature. The two best-known methods are the following:

$H(z) = P(0 < Z < z)$ and
$\Phi(z) = P(Z \leq z)$, for positive values of z

$\Phi(z)$ is the cumulative distribution function of the n(0, 1) distribution, and as $\Phi(0) = 0.5$, it follows directly that $\Phi(z) = 0.5 + H(z)$ for $z > 0$.

In Table B.2 of Appendix B $H(z)$ is tabulated for z from 0.00 to 4.00 in steps of 0.01. According to this table the last required probability of Example 10.15 is therefore

$$P(0 < Z < 1.67) = 0.45254$$

Figure 10.10 contains a graphic presentation of the process of standardization and the use of the normal tables for Example 10.15. From this figure we immediately note that

$$\begin{aligned} P(20 < X < 25) &= P(-1.67 < Z < 0) \\ &= P(0 < Z < 1.67) \\ &= 0.45254 \end{aligned}$$

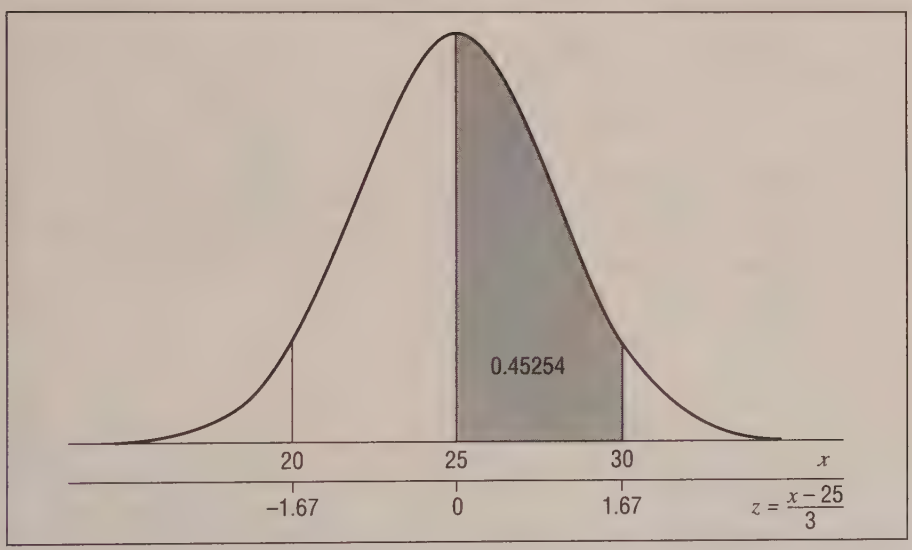


Figure 10.10 Standardization and the use of normal tables

This result is derived from the fact that the normal curve is symmetrical around the mean μ, and this is why the function $H(z)$ defined above is tabulated only for positive values of z.

From the probability interpretation of the area below the density function it is obvious that the total area below the curve has to be equal to 1. Due to symmetry it therefore follows that the areas to the right and to the left of $z = 0$ are 0.5. Using these characteristics of the normal curve it is possible to determine any probability expression for a normal distribution with known mean μ and variance σ^2. The following example illustrates the use of the tables.

EXAMPLE 10.16

Suppose X has a n(10, 16) distribution. The following problems are typical:

(a) $P(12 < X < 20)$ (b) $P(4 < X < 8)$
(c) $P(X > 15)$ (d) $P(X < 5)$
(e) $P(6 < X < 18)$ (f) b so that $P(X < b) = 0.25$

The determination of these with $\mu = 10$ and $\sigma = 4$ is as follows:

(a) $P(12 < X < 20)$

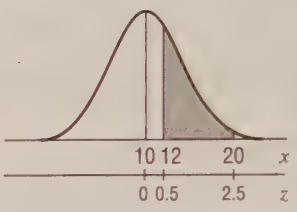
$$= P\left(\frac{12 - 10}{4} < \frac{X - 10}{4} < \frac{20 - 10}{4}\right)$$

$$= P(0.5 < Z < 2.5)$$

$$= H(2.5) - H(0.5)$$

$$= 0.49379 - 0.19146$$

$$= 0.30233$$



(b) $P(4 < X < 8)$

$$= P\left(\frac{4 - 10}{4} < \frac{X - 10}{4} < \frac{8 - 10}{4}\right)$$

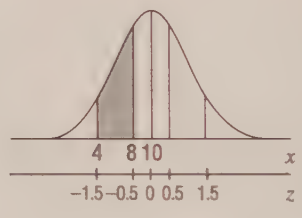
$$= P(-1.5 < Z < -0.5)$$

$$= P(0.5 < Z < 1.5)$$

$$= H(1.5) - H(0.5)$$

$$= 0.43319 - 0.19146$$

$$= 0.24173$$



(c) $P(X > 15)$

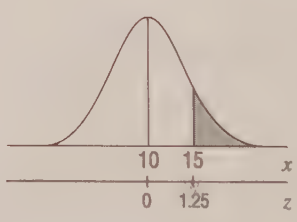
$$= P\left(\frac{X - 10}{4} > \frac{15 - 10}{4}\right)$$

$$= P(Z > 1.25)$$

$$= 0.5 - H(1.25)$$

$$= 0.5 - 0.39435$$

$$= 0.10565$$



(d) $P(X < 5)$

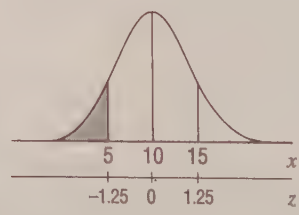
$$= P\left(\frac{X - 10}{4} < \frac{5 - 10}{4}\right)$$

$$= P(Z < -1.25)$$

$$= 0.5 - H(1.25)$$

$$= 0.5 - 0.39435$$

$$= 0.10565$$



(e) $P(6 < X < 18)$

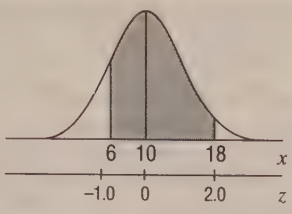
$$= P\left(\frac{6-10}{4} < \frac{X-10}{4} < \frac{18-10}{4}\right)$$

$$= P(-1.0 < Z < 2.0)$$

$$= H(1.0) + H(2.0)$$

$$= 0.34134 + 0.47725$$

$$= 0.81859$$



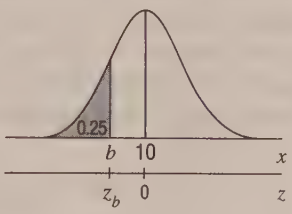
(f) $P(X < b)$

$$= P\left(\frac{X-10}{4} < \frac{b-10}{4}\right)$$

$$= P(Z < z_b)$$

$$= \Phi(z_b)$$

$$= 0.25$$



On approximation, therefore

$$z_b = -0.67$$

$$= \left(\frac{b-10}{4}\right)$$

so that $b \approx 4(-0.67) + 10$

$$= 7.32$$

Notes

(a) Example 10.16 clearly shows that a sketch is indispensable when the normal tables are used.

(b) As seen from the solution to Example 10.16 (f), the normal tables can also be used to determine the value of z for which $\Phi(z) = p$, with p a specified probability. Symbolically this can be written as follows:

> $\Phi(z) = p$ implies that
>
> $z = \Phi^{-1}(p)$
>
> with Φ^{-1} the so-called **inverse distribution function** of the n(0, 1) distribution.

In addition to the above method, the value of $\Phi^{-1}(p)$ for $0 < p < 1$ can also be determined directly from tables of the inverse normal distribution function. Table B.14 in Appendix B contains the values of the inverse

normal distribution function $\Phi^{-1}(p)$ for $p \geq 0.5$. For $p < 0.5$, $\Phi^{-1}(p)$ can be determined from the following relationship:

$$\Phi^{-1}(p) = -\Phi^{-1}(1 - p)$$

For the above example the value of z had to be determined, for which $\Phi(z) = 0.25$. We can read, from Table B.14,

$$z = \Phi^{-1}(0.25) = -\Phi^{-1}(0.75) = -0.674$$

which is more accurate than the value found in Example 10.16(f) (see also Note (c) below).

(c) In some cases it is not possible to determine the exact $H(z)$ value for a given z value (or vice versa). In such a case the value can be approximated, as in Example 10.16(f), or the technique of **linear interpolation** can be used to determine a more accurate value. In the case of the above example, linear interpolation would have led to the following result:

$$z_b = -\left\{0.67 + \frac{0.25 - 0.24857}{0.25175 - 0.24857} \times (0.68 - 0.67)\right\}$$

$$= -0.675$$

so that $b = 4(-0.675) + 10$

$$= 7.30$$

The small difference between the z value of Note (b) above, namely -0.674, and the z value determined by means of linear interpolation, namely -0.675, is due to the fact that linear interpolation leads to only approximate answers.

EXAMPLE 10.17

In Example 10.2 we calculated the relative frequency distribution of the price per square metre of a population of 843 houses. We noted, *inter alia*, that the histogram and frequency polygon of this distribution can be approximated by the smooth curve in Figure 10.3. This curve is also bell shaped but somewhat positively skewed. Since the exact mathematical form of the density function concerned is unknown, it might be of value to assume that the distribution concerned can be approximated by a normal distribution. It can be shown that the mean and standard deviation of the 843 prices are given by

$$\mu = 922 \text{ and } \sigma = 349 \text{ rand}$$

It therefore follows, for example, that

$$P(100 < X < 300) = P\left(\frac{100 - 922}{349} < \frac{X - 922}{349} < \frac{300 - 922}{349}\right)$$

$$= P(-2.36 < Z < -1.78)$$

$$= H(2.36) - H(1.78)$$

$$= 0.0284$$

This value compares well with the corresponding relative frequency of 0.0202, as given in Table 10.3. Table 10.6 provides a comparison of the relative frequencies of Table 10.3 and the corresponding probabilities as calculated for the $n(922, (349)^2)$ distribution.

Table 10.6 Comparison of relative frequencies and probabilities obtained from the normal approximation of the prices per square metre of 843 Johannesburg residences

Class interval	Relative frequency	Normal probability
[100, 300)	0.0202	0.0284
[300, 500)	0.0854	0.0756
[500, 700)	0.1851	0.1480
[700, 900)	0.2064	0.2150
[900, 1 100)	0.2052	0.2189
[1 100, 1 300)	0.1435	0.1650
[1 300, 1 500)	0.0973	0.0916
[1 500, 1 700)	0.0415	0.0360
[1 700, 1 900)	0.0107	0.0103
[1 900, 2 100)	0.0047	0.0022
	1.0000	0.9910

Note that X is a $n(922, (349)^2)$ variable, and that $P(100 < X < 2\,100) = 0.9910$, while theoretically for any normal distribution $P(-\infty < X < \infty) = 1.0$.

Although the normal approach in this example can be regarded as relatively adequate, the approximation cannot be assessed without using objective tests. There are various so-called **goodness-of-fit tests** (see for instance Chapter 15), but a simple graphic procedure known as the **normal probability plot** will suffice here:

Let $x_1, x_2, \ldots, x_n$ be a random sample of observations of a stochastic variable X, and let $\bar{x}$ and s be the arithmetic mean and standard deviation of the sample. Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ denote the array of the x values and $z_{(1)}, z_{(2)}, \ldots, z_{(n)}$ the standardized values of the array with $z_{(i)} = (x_{(i)} - \bar{x})/s$, $i = 1, 2, \ldots, n$. Define $y_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$, where Φ^{-1} is the inverse distribution function of the $n(0, 1)$ distribution, $i = 1, 2, \ldots, n$. If the points $(y_1, z_{(1)}), (y_2, z_{(2)}), \ldots, (y_n, z_{(n)})$ are plotted graphically and lie more or less on a straight line with intercept 0 and slope 45°, we can assume with a reasonable degree of certainty that the stochastic variable X approximately has a $n(\bar{x}, s^2)$ distribution.

The normal probability plot is illustrated in Example 10.18.

EXAMPLE 10.18

A class of 46 students in a Mathematical Statistics course consists of two groups: Group 1 comprises 20 commerce students and Group 2 26 natural science students. The means and standard deviations of the two groups and the joint group are given in the accompanying table.

	Group 1	Group 2	Combined
n	20	26	46
$\bar{x}$	66.9	58.7	62.3
s	14.4	13.0	14.1

Table 10.7, column 2, contains an array of the final marks, $x_{(i)}$, that Group 2 scored in the course. The third column contains the standardized values $z_{(i)} = \frac{x_{(i)} - \bar{x}}{s}$ and the last column the inverse standard normal values $y_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$. Note that $\Phi^{-1}\left(\frac{i}{n+1}\right) = -\Phi^{-1}\left(1 - \frac{i}{n+1}\right)$. Similar tables can be compiled for Group 1 and the joint class. In Figure 10.11(a), (b) and (c) the points $(y_1, z_{(1)})$, $(y_2, z_{(2)})$, ..., $(y_n, z_{(n)})$ were plotted for the three cases. The **convex** form of the scatter diagram in (a) indicates a slightly negatively skewed distribution (that is, not normal); while the **concave** form of the scatter diagram in (b) indicates a positively skewed (not normal) distribution. The straight-line tendency of the scatter diagram in (c) signifies a normal distribution.

Table 10.7 Calculation table for testing normality

i	$x_{(i)}$	$z_{(i)} = \frac{x_{(i)} - \bar{x}}{s}$	$y_i = \Phi^{-1}\left(\frac{i}{27}\right)$
1	41	–1.362	–1.787
2	43	–1.208	–1.447
3	43	–1.208	–1.221
4	47	–0.900	–1.045
5	47	–0.900	–0.896
6	48	–0.823	–0.765
7	49	–0.746	–0.646
8	50	–0.669	–0.536
9	51	–0.592	–0.432
10	52	–0.515	–0.332
11	53	–0.438	–0.235
12	53	–0.438	–0.141
13	55	–0.285	–0.048
14	55	–0.285	0.048
15	58	–0.054	0.141
16	60	0.100	0.235
17	61	0.177	0.332
18	61	0.177	0.432
19	63	0.331	0.536
20	69	0.792	0.646
21	69	0.792	0.765
22	71	0.946	0.896
23	75	1.254	1.045
24	79	1.562	1.221
25	82	1.792	1.447
26	91	2.485	1.787

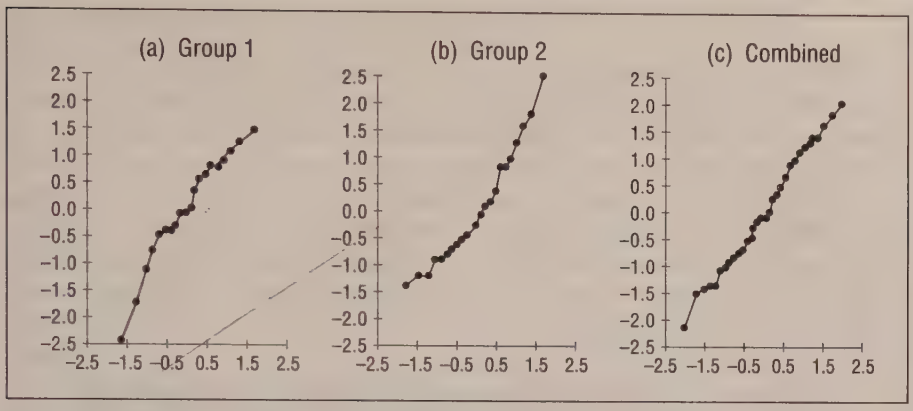


Figure 10.11 Normal probability plot of data

8 Sampling distributions and the central limit theorem

In Chapter 4, Section 2, we mentioned that the arithmetic means of samples of the same size vary from one sample to the next. If a large number of random samples of, say, n observations are drawn from the same population, and $\bar{x}$ is calculated each time, these values of $\bar{x}$ can be grouped in a frequency table. The distribution of the $\bar{x}$ values can then be represented graphically by a histogram or frequency polygon. Descriptive measures such as $\bar{x}_{\bar{x}}$ and $s_{\bar{x}}$, the arithmetic mean and standard deviation of the distribution of the $\bar{x}$ values, can also be calculated. Consider the following example that is given as an explanation.

EXAMPLE 10.19

A manufacturer of electronic components packs a certain product in containers that hold 120 identical articles. A potentially large consumer of the components requires information on the life expectancy of the component. Among other things he wants to know what percentage of containers holds components of which the **mean** life is shorter than eighty days. The manufacturer has just received records on the life of the 120 components from one container and the data is represented in the histogram of Figure 10.12(a). This histogram corresponds with information the manufacturer has on X = life of component, namely that X has an exponential distribution with an average of 100 days. According to the characteristics of the exponential distribution (see Section 6) it would seem that about 55% of all components have a life expectancy of less than eighty days. However, the customer needs to know what percentage of containers' components have an **average** life expectancy of less than eighty days.

The problem can be **simulated** by taking a large number of samples of a stochastic variable which has an exponential distribution with a mean of $\mu = 100$. This process was carried out with the aid of a computer. The arithmetic means $\bar{x}_i$ of

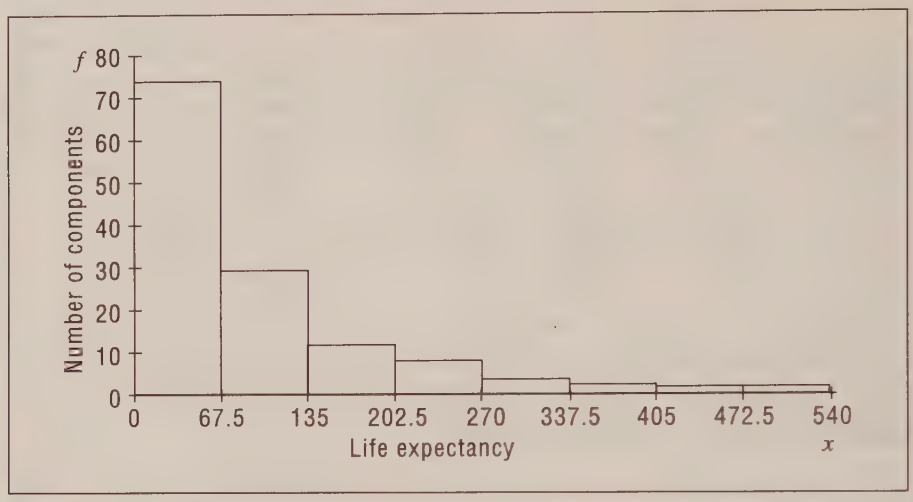


Figure 10.12(a) Histogram of life expectancy of 120 electronic components

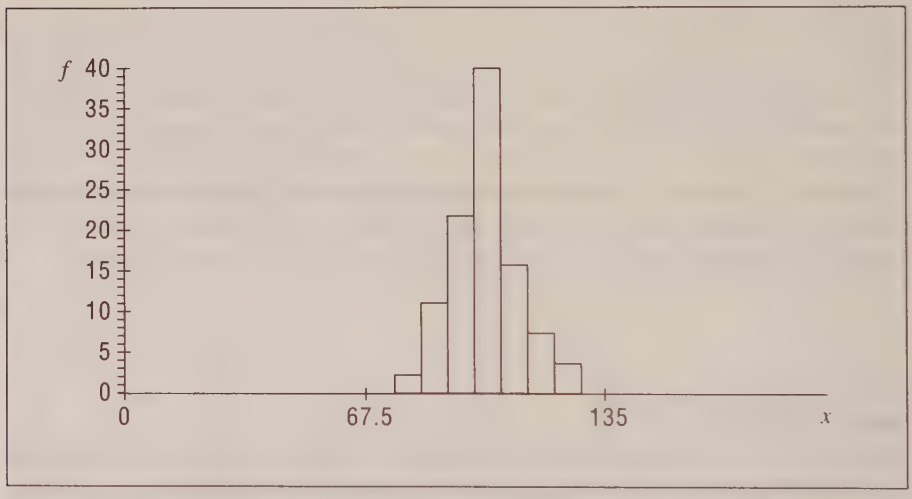


Figure 10.12(b) Histogram of arithmetic means of 100 random samples of size 120

100 random samples of size 120 each were determined and represented graphically in the histogram in Figure 10.12(b).

Since approximately 1.0% of the total area of the histogram in Figure 10.12(b) lies below the value of 80, the answer to the customer's question is therefore: The mean life expectancy of components in only 1% of all containers is shorter than eighty days.

For the distribution of Figure 10.12(a) it was also found that:

$\bar{x} = 102.7$ and $s^2 = 10\,371.0$

while for the distribution in Figure 10.12(b), it was found that:

$\bar{x}_{\bar{x}} = 100.5$ and $s^2_{\bar{x}} = 77.5$

8.1 Statistics and sampling distributions

One of the most important observations that can be deduced from the above example is the fact that, in repeated sampling, the arithmetic mean of a sample has a distribution for which a histogram like the one in Figure 10.12(b) can be drawn. This implies that the arithmetic mean for **all possible samples of the same size** can be regarded as a stochastic variable, namely $\bar{X}$. Similarly, we can calculate the standard deviation s for each sample and draw a histogram of the distribution of the s values. The standard deviation of a sample from a population can therefore also be regarded as a stochastic variable S. In accordance with this view, it is customary to refer to the stochastic variables $\bar{X}$ and S as **statistics**.

If all samples of size n are drawn from a population and the value of some or other statistic is calculated for each sample, all the values have a certain distribution known as the **sampling distribution** of the statistic.

To study this concept of 'sampling distribution' consistently, we use the following notation when drawing a simple random sample from a population of x values:

$X_1, X_2, \ldots, X_n$ respectively represent the first, second, ..., nth value of the random sample, while $x_1, x_2, \ldots, x_n$ respectively represent the observed values of the stochastic variables and are called the observed random sample.

In the above notation we can now define, for example, the following statistics for the random sample $X_1, X_2, \ldots, X_n$:

$$\bar{X} = \frac{\sum X}{n} \quad \text{and} \quad S = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$$

while $\bar{x} = \frac{\sum x}{n}$ and $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$

are the observed values of the stochastic variables $\bar{X}$ and S and are based on the observed random sample $x_1, x_2, \ldots, x_n$.

If $\bar{X}$ is seen as a stochastic variable, it follows that the histogram of $\bar{x}$ values in Figure 10.12(b) will become a smooth curve if the number of samples are greatly increased to, for instance, 1 000 or more. This smooth curve is obviously the density function of the distribution of $\bar{X}$.

A comparison of the two histograms in Figure 10.12 and of $\bar{x}$ and s^2 with $\bar{x}_{\bar{x}}$ and $s^2_{\bar{x}}$ leads to the following interesting observations:

(a) The $\bar{x}$ distribution (case (b)) is far more symmetrical than the x distribution (case (a)) and has the typical normal shape.

(b) The arithmetic means of the two distributions are approximately the same ($\bar{x} = 102.7$ and $\bar{x}_{\bar{x}} = 100.5$).

(c) The variance $s^2_{\bar{x}} = 77.5$ of the $\bar{x}$ distribution is much smaller than the variance $s^2 = 10\,371.0$ of the x distribution.

The above empirical conclusions illustrate two important results. The first of these can be formulated as follows:

If $\bar{X}$ is the arithmetic mean of a sample of size n from a population with mean μ and variance σ^2, the mean and variance of the sampling distribution of $\bar{X}$ are given by

$$\mu_{\bar{X}} = \mu \quad \text{and} \quad \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

For the data in Example 10.19, we can then say that

$$\mu_{\bar{X}} = \mu = 100$$

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

$$= \frac{10\,000}{120}$$

$$= 83.3$$

since it is known that, for the exponential distribution, $\sigma^2 = \mu^2$ (see Section 6).

8.2 The central limit theorem

We now consider the shape of the density function of $\bar{X}$. To begin, study the following simple experiment:

(a) One dice is rolled and the value of X, the outcome, is recorded. This process is repeated many times. The frequency distribution of x values is represented graphically by means of a bar chart.

(b) Two dice are rolled simultaneously and the value of $\bar{X}$, the mean outcome of rolling the two dice, is recorded. The process is repeated many times. A bar chart is drawn up for the $\bar{x}$ values.

(c) Repeat (b) for 3, 4, ... dice.

From this experiment we note that although the outcome of rolling one dice is distributed discreetly with $P(X = 1) = P(X = 2) = \ldots = P(X = 6) = \frac{1}{6}$, the bar chart of $\bar{X}$ tends increasingly towards a normal curve as n, the number of dice, increases.

This experiment and the one discussed in Example 10.19, illustrates the second important result referred to, namely the **central limit theorem**:

The arithmetic mean of a random sample of size n from **any** population with finite variance σ^2, approximately has a $n(\mu, \frac{\sigma^2}{n})$ distribution, provided n is fairly large.

It should be emphasized that the central limit theorem applies to sampling from **any** population with finite variance: In Example 10.19 samples of the continuous but positively skewed exponential distribution are discussed, while in the above experiment samples of a discrete stochastic variable are investigated. The value of n for which the normal approximation is satisfactory is mainly determined by the degree of skewness and kurtosis of the underlying density function. A sample size of $n = 30$ is usually regarded as adequate for a satisfactory normal approximation of the sampling distribution of $\bar{X}$.

EXAMPLE 10.20

Study the distribution of $\bar{X}$, the mean life expectancy of a sample of 120 electronic components. It is known that the mean and standard deviation of the life expectancy of all components are given by $\mu = 100$ days and $\sigma = 100$ days (see Example 10.19). Determine

(a) $P(\bar{X} < 80)$;

(b) $P(90 < \bar{X} < 110)$;

(c) The value b for which $P(\bar{X} > b) = 0.10$

From the central limit theorem it now follows that $\bar{X}$ is approximately distributed $n(100; \frac{100^2}{120})$. Therefore

$$\text{(a)} \quad P(\bar{X} < 80) = P\left(\frac{\bar{X} - 100}{\frac{100}{\sqrt{120}}} < \frac{80 - 100}{\frac{100}{\sqrt{120}}}\right)$$

$$= P(Z < -2.19)$$

$$= 0.5 - H(2.19)$$

$$= 0.5 - 0.48574$$

$$= 0.01426$$

$$\approx 0.01$$

$$\text{(b)} \quad P(90 < \bar{X} < 110) = P\left(\frac{90 - 100}{\frac{100}{\sqrt{120}}} < \frac{\bar{X} - 100}{\frac{100}{\sqrt{120}}} < \frac{110 - 100}{\frac{100}{\sqrt{120}}}\right)$$

$$= P(-1.10 < Z < 1.10)$$

$$= 2 \times H(1.10)$$

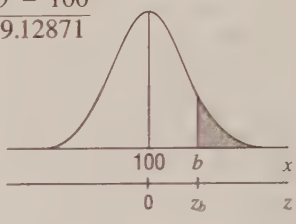
$$= 2 \times 0.3643$$

$$= 0.7286$$

(c) $P(\bar{X} > b) = 0.10$

implies that $P\left(\dfrac{\bar{X} - 100}{\frac{100}{\sqrt{120}}} > \dfrac{b - 100}{\frac{100}{\sqrt{120}}}\right) = 0.10$

Set $z_b = \dfrac{b - 100}{\frac{100}{\sqrt{120}}} = \dfrac{b - 100}{9.12871}$



it then follows that $P(Z < z_b) = 0.90$

so that $H(z_b) = 0.40.$

Therefore $z_b = 1.28$

so that $b = 100 + 1.28 \times 9.12871$

≈ 111.7 days

Notes

(a) From the central limit theorem it can be directly deduced that $\dfrac{(\bar{X} - \mu)}{\sigma \sqrt{n}}$ approximately has the n(0, 1) distribution. This expression is known as a **pivotal quantity** as it is actually the pivot around which the statistical inference regarding μ revolves if σ is known (see Chapters 11 and 12).

(b) After mathematical manipulation of this pivotal quantity we find that

$$\frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} = \frac{\sum X - n\mu}{\sqrt{n}\,\sigma}$$

From this a more general formulation of the central limit theorem follows, namely:

> If $X_1, X_2, \ldots, X_n$ is a random sample from a population with mean μ and finite variance σ^2, $\sum X$ will approximately be distributed $n(n\mu, n\sigma^2)$, provided n is fairly large.

(c) If $X_1, X_2, \ldots, X_n$ is a random sample from a $n(\mu, \sigma^2)$ population, $\bar{X}$ is distributed $n(\mu, \frac{\sigma^2}{n})$. This is also a very important result from which it follows that the above pivotal quantity Z is, in this case, distributed exactly n(0, 1).

8.3 The normal approximation of the binomial distribution

Consider a binomial experiment as explained in Section 3. Here we could also have defined the number of successes, X, in n independent trials in terms of a simple random sample $X_1, X_2, \ldots, X_n$ as follows:

$$X = \Sigma X_i$$

where

$$X_i = \begin{cases} 1 \text{ for a success in the } i\text{th trial} \\ 0 \text{ for a failure in the } i\text{th trial} \end{cases}$$

with

$$P(X_i = 1) = \pi \text{ and } P(X_i = 0) = 1 - \pi$$

The mean and standard deviation of X_i can be calculated as follows (see Section 2.3):

$$\begin{aligned} \mu_{X_i} &= (1)\pi + (0)(1 - \pi) \\ &= \pi \\ \sigma_{X_i} &= \sqrt{(1)^2 \pi + (0)^2 (1 - \pi) - \mu_{X_i}^2} \\ &= \sqrt{\pi - \pi^2} \\ &= \sqrt{\pi(1 - \pi)} \end{aligned}$$

Using this representation of the binomial stochastic variable and the more general formulation of the central limit theorem given in Note (b) above, we obtain the following important result:

If X has a bi(n, π) distribution, X will approximately have a n($n\pi$, $n\pi(1 - \pi)$) distribution for large values of n, while $P = \frac{X}{n}$ approximately has a $\text{n}\left(\pi, \frac{\pi(1 - \pi)}{n}\right)$ distribution.

This result enables us to determine approximate probability expressions of the binomial distribution by means of the normal distribution. The success of this normal approximation depends on the general form of the underlying binomial distribution. The more skewed the binomial distribution, the larger the value of n required for a satisfactory approximation. A general guideline is that $n \geq 30$ if π is more or less 0.5, while n must progressively be larger the more n differs from 0.5.

EXAMPLE 10.21

Suppose X has a bi(40, 0.5) distribution. In this case

$$n\pi = 40 \times 0.5 = 20 \quad \text{and} \quad n\pi(1 - \pi) = 40 \times 0.5 \times 0.5 = 10$$

Now
$$\begin{aligned} P(X \leq 15) &= P\left(\frac{X - 20}{\sqrt{10}} \leq \frac{15 - 20}{\sqrt{10}}\right) \\ &= P(Z \leq -1.58) \end{aligned}$$

$$= 0.5 - H(1.58)$$
$$= 0.5 - 0.44295$$
$$= 0.05705$$

(The exact value of this probability is 0.0769.)

Similarly $P(16 \leq X \leq 28) = P\left(\frac{16-20}{\sqrt{10}} \leq \frac{X-20}{\sqrt{10}} \leq \frac{28-20}{\sqrt{10}}\right)$

$$= P(-1.26 \leq Z \leq 2.53)$$
$$= H(1.26) + H(2.53)$$
$$= 0.39617 + 0.49430$$
$$= 0.89047$$

(The exact value of this probability is 0.9199.)

Finally $P(X = 20) = P(X \leq 20) - P(X \leq 19)$

$$= P(Z \leq 0) - P(Z \leq -0.32)$$
$$= 0.5 - \{0.5 - H(0.32)\}$$
$$= 0.12552$$

(The exact value of this probability is 0.1254.)

Notes

(a) The expression

$$\frac{P - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

is the **pivotal quantity** around which the statistical inference with respect to the parameter π revolves (see Chapters 11 and 12).

(b) In the normal approximation of the binomial distribution remember that the binomial distribution is a discrete distribution and the normal distribution is a continuous distribution. If n is not very large, the approximation can be improved by applying the so-called **continuity correction**.

Figure 10.13 can be used to illustrate the following discussion of this correction. The probability of the binomial stochastic variable X lying between two integers a and b (both included), is given by the sum of the probabilities of the values a, $a + 1$, ..., b. Since these probabilities are represented by the heights of the rectangles in the histogram and the base of each rectangle is equal to 1, the sum of the probabilities of the values between a and b is therefore equal to the area of the histogram from $x = a - \frac{1}{2}$ to $b + \frac{1}{2}$. We can thus say that

$$P(a \le X \le b) = P\left(\frac{a - \frac{1}{2} - n\pi}{\sqrt{n\pi(1-\pi)}} \le \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}} \le \frac{b + \frac{1}{2} - n\pi}{\sqrt{n\pi(1-\pi)}}\right)$$

$$\approx P\left(Z \le \frac{b + \frac{1}{2} - n\pi}{\sqrt{n\pi(1-\pi)}}\right) - P\left(Z \le \frac{a - \frac{1}{2} - n\pi}{\sqrt{n\pi(1-\pi)}}\right)$$

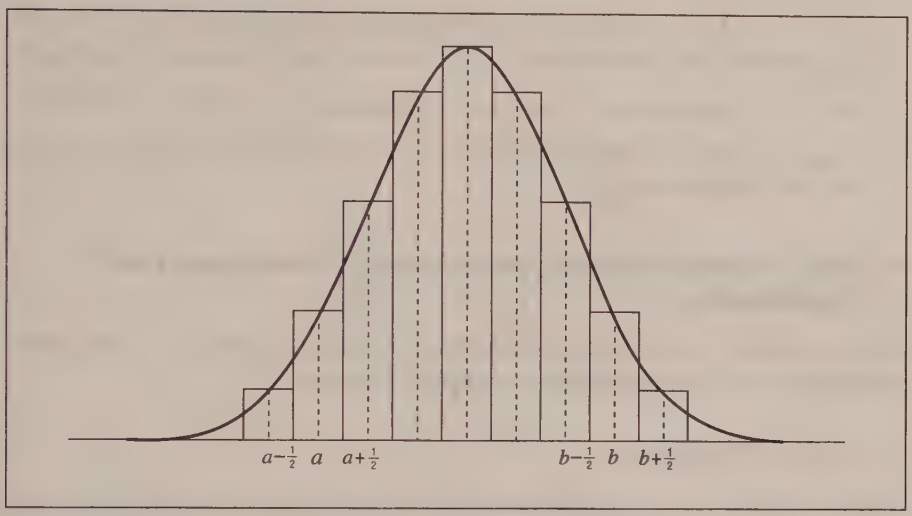


Figure 10.13 Normal approximation of the binomial distribution

For the approximate determination of $P(X \le c)$ and $P(X \ge d)$ the following relations should be used:

$$P(X \le c) = P(0 \le X \le c)$$

and

$$P(X \ge d) = P(d \le X \le n)$$

The continuity correction is then applied as for a and b above.

We can, for example, obtain a better approximation of $P(16 \le X \le 28)$, as calculated in Example 10.21, by

$$\begin{aligned} P(16 \le X \le 28) &= P\left(Z \le \frac{28.50 - 20}{\sqrt{10}}\right) - P\left(Z \le \frac{15.5 - 20}{\sqrt{10}}\right) \\ &= P(Z \le 2.69) - P(Z \le -1.42) \\ &= H(1.42) + H(2.69) \\ &= 0.42220 + 0.49643 \\ &= 0.91863 \end{aligned}$$

This value is very close to the exact value of 0.9199.

If we work with $P = \frac{X}{n}$ the application of the continuity correction yields the following expression:

$$P\left(c \leq \frac{X}{n} \leq d\right) \approx P\left(Z \leq \frac{d + \frac{1}{2n} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}\right) - P\left(Z \leq \frac{c - \frac{1}{2n} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}\right)$$

(c) If X has a po(λ) distribution, probability expressions for X can be approximated in a similar way to the one used for the binomial distribution, as $\frac{X-\lambda}{\sqrt{\lambda}}$ approximately has a n(0, 1) distribution provided λ is relatively large. Usually a satisfactory normal approximation is obtained for the Poisson distribution if $\lambda > 10$.

9 The χ^2 distribution, the t distribution and the F distribution

In this section three more continuous stochastic variables are discussed, particularly for their importance to statistical inference.

9.1 The χ^2 distribution

The χ^2 variable can assume only positive real values and the density function is positively skewed. The mathematical function $f(v)$ that is associated with this curve depends on one parameter, the so-called degrees of freedom v and we then refer to the $\chi^2(v)$ distribution.

The mean and standard deviation of the $\chi^2(v)$ distribution are given by

$\mu = v$ and $\sigma = \sqrt{2v}$

Figure 10.14 illustrates the density function of the $\chi^2(9)$ distribution.

The following result is of importance for the χ^2 distribution:

> If $X_1, X_2, \ldots, X_n$ is a random sample from a n(μ, σ^2) population, the sample distribution of $Y = \frac{(n-1)S^2}{\sigma^2}$ is that of the $\chi^2(n-1)$ distribution.

From this result it follows, for instance, that if the variance of a sample of size 10 from a n(25, 16) distribution is given by $s^2 = 14.6$, then $y = \frac{9 \times 14.6}{16} \approx 8.2$ is an observation of a $\chi^2(9)$ variable.

9.1.1 Tables of the χ^2 distribution

In contrast to the normal distribution, a single table of the χ^2 distribution cannot be drawn up to determine the probability expressions. A separate table should be compiled for each value of the parameter v. Since certain sections of

these tables are used more often than others, only those sections for different values of v are summarized in one table, as given in Table B.4 in Appendix B. In this table χ^2_v denotes a stochastic variable that has the $\chi^2(v)$ distribution. The tabulated values are indicated by $\chi^2_{v,\alpha}$ where

$$P(\chi^2_v \geq \chi^2_{v,\alpha}) = \alpha$$

with α a prespecified value.

Figure 10.14 illustrates this notation for the case where

$v = 9$, $\alpha = 0.05$ and $\chi^2_{9,0.05} = 16.919$

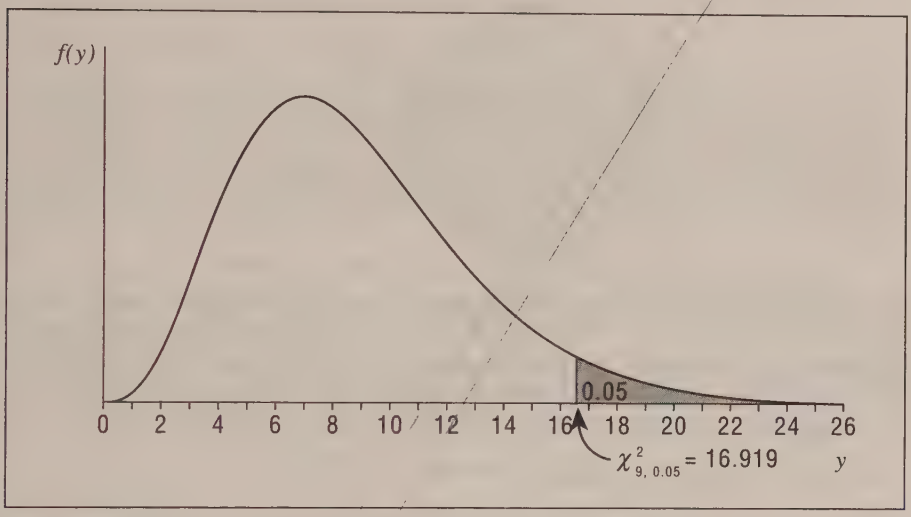


Figure 10.14 Illustration of the use of χ^2 tables

EXAMPLE 10.22

Suppose we need to determine the value of

(a) $\chi^2_{24,0.025}$ and

(b) $P(\chi^2_{24} \geq 35)$

(a) From Table B.4 $\chi^2_{24,0.025} = 39.364$

(b) From Table B.4 $\chi^2_{24,0.05} = 36.415$

$\chi^2_{24,0.10} = 33.196$

so we can only say that

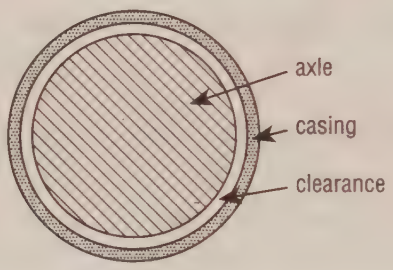
$$0.05 < P(\chi^2_{24} \geq 35) < 0.10$$

More detailed tables are therefore required if this probability is to be determined precisely.

The next example is a practical illustration of the use of the above concepts.

EXAMPLE 10.23

An engineering firm that manufactures parts for machines experiences problems with the variation in the cross-section of a certain axle casing. According to specifications the cross-section should be 800 mm. The manufacturing process can produce axle casings for which the cross-section has a n(800, 0.25) distribution. If the variance becomes too great, problems are experienced with friction and wear and tear. Where there is too great a variation in the cross-section, an unacceptably high proportion of axles fit either too tightly or too loosely in the axle casing (see accompanying sketch).



A random sample of twelve axles was measured and the following measures for the cross-section were found: $\bar{x} = 800.01$, $s^2 = 0.3975$. According to the supervisor of the process so high a value of s^2 is quite possible, even if the population variance is $\sigma^2 = 0.25$. The question then is whether or not the supervisor's assertion is correct.

Using the χ^2 distribution the problem can be solved as follows:

$$P\left(S^2 \geq 0.3975 \mid \sigma^2 = 0.25\right)$$

$$= P\left(\frac{(n-1)S^2}{\sigma^2} \geq \frac{(n-1)0.3975}{\sigma^2} \mid \sigma^2 = 0.25\right)$$

$$= P\left(\chi^2_{11} \geq 17.49\right) \text{ by replacing } n = 12 \text{ and } \sigma^2 = 0.25.$$

From Table B.4 in Appendix B it follows that

$$0.05 < P\left(\chi^2_{11} \geq 17.49\right) < 0.10$$

We can therefore say that the probability of obtaining a sample variance of 0.3975 or higher is between 5% and 10%, that is, very low. We can thus confidently conclude that the supervisor's assertion is incorrect.

Notes

(a) The quantity $Y = \dfrac{(n-1)S^2}{\sigma^2}$ is the **pivotal quantity** on which the statistical inference with respect to σ^2 is based (see Chapters 11 and 12). The above example is an illustration of this.

(b) It can be proved that the standardized stochastic variable $\frac{Y - \nu}{\sqrt{2\nu}}$ approximately has a n(0, 1) distribution, provided ν is greater than 100.

9.2 The t distribution

Under the pseudonym 'Student', W.S. Gosset published an article in 1908 in which the characteristics of the so-called t distribution were discussed for the first time. For this reason it is often called 'Student's t distribution'. As with the χ^2 distribution, the mathematical function $f(t)$ that is associated with the t curve depends on a single parameter ν called the number of degrees of freedom of the t distribution. We also refer to the t(ν) distribution.

The following result is of primary interest here:

If $X_1, X_2, \ldots, X_n$ is a random sample from a n(μ, σ^2) population, the sample distribution of $T = \frac{(\bar{X} - \mu)}{S/\sqrt{n}}$ is that of the t($n - 1$) distribution.

From this result it follows, for instance, that if a sample of size 10 is drawn from a n(25, σ^2) population and $\bar{x} = 24.6$, $s = 3.73$, then $t = \frac{(24.6 - 25.0)}{3.73/\sqrt{10}}$ is an observation of a t(9) variable.

It is important to note that for $X_1, X_2, \ldots, X_n$ a random sample from a n(μ, σ^2) population

$Z = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}}$ has a n(0, 1) distribution and

$T = \frac{(\bar{X} - \mu)}{S/\sqrt{n}}$ has a t(n – 1) distribution.

Since s for an observed random sample should not differ too much from σ, we can expect the density functions of Z and T to be fairly similar, particularly for large values of n. This aspect is clearly illustrated in Figure 10.15, where the density functions of the t(4) and n(0, 1) distributions are presented graphically. The density function of the t distribution is symmetrically bell shaped around zero, with tail sections higher than those of the n(0, 1) distribution. The importance of the differences between the n(0, 1) and the t distributions follows from the fact that in statistical inference the areas below the tail sections of the density functions are important: Compare, for example, the areas below the two curves in Figure 10.15 for $x > 3$. As the degrees of freedom ν increases, the density function of the t(ν) distribution tends more towards the curve of the n(0, 1) distribution.

The mean and standard deviations of the t(ν) distribution are respectively given by $\mu = 0$ and $\sigma = \sqrt{\frac{\nu}{\nu - 2}}$, provided $\nu > 2$.

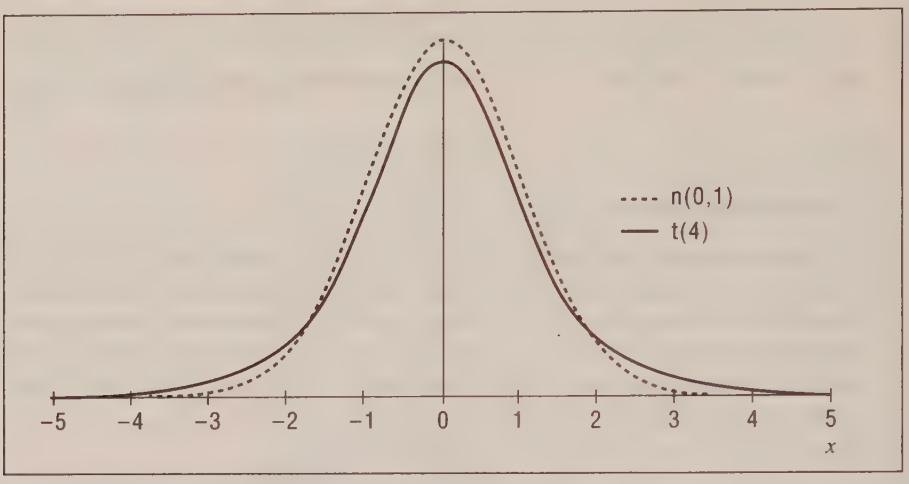


Figure 10.15 Density functions of the t(4) and n(0, 1) distributions

9.2.1 Tables of the t distribution

As with the χ^2 distribution the usual tables of the t distribution contain only the values of $t_{\nu,\alpha}$ for which

$$P(T_\nu \geq t_{\nu,\alpha}) = \alpha$$

where T_ν is a stochastic variable that has a t(ν) distribution. Since the t distribution is symmetrical around zero, only the positive values of $t_{\nu,\alpha}$ need to be tabulated. These values appear in Table B.3 in Appendix B. Figure 10.16 illustrates the interpretation of the tables. For $\nu = 9$ and $\alpha = 0.05$ it follows that

$$t_{9,0.05} = 1.833 \text{ and } -t_{9,0.05} = -1.833.$$

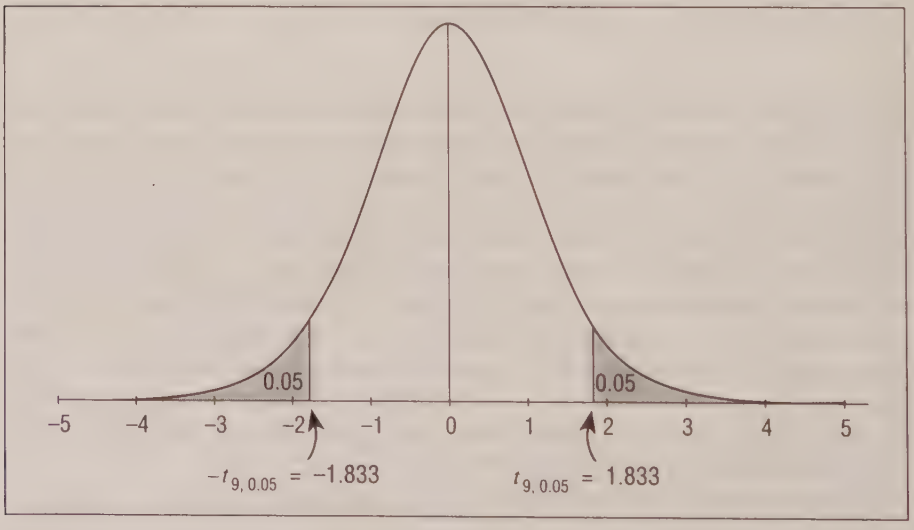


Figure 10.16 Illustration of the use of t tables

EXAMPLE 10.24

Random samples of size 25 are drawn from a $n(\mu, \sigma^2)$ population.

(a) For which percentage of the samples will $\frac{(\bar{X} - \mu)}{S/\sqrt{n}}$ be greater than 2.064?

(b) Determine b so that $P(|T_{24}| < b) = 0.90$.

Using the above theory, it follows that

(a) $$P\left(\frac{(\bar{X} - \mu)}{S/\sqrt{n}} > 2.064\right)$$
$$= P(T_{24} > 2.064)$$
$$= 0.025, \text{ using Table B.3 in Appendix B.}$$

Therefore: for 2.5% of the samples $\frac{(\bar{X} - \mu)}{S/\sqrt{n}}$ is greater than 2.064.

(b) $$P(|T_{24}| < b)$$
$$= P(-b < T_{24} < b)$$
$$= 2P(0 < T_{24} < b) \quad \text{from the symmetry of the t distribution}$$
$$= 0.90$$

Therefore $P(T_{24} \geq b) = \frac{1}{2}(1 - 0.90)$
$$= 0.05$$

from which it can be derived, with the aid of Table B.3, that $b = 1.711$.

Notes

(a) The quantity $T = \frac{(\bar{X} - \mu)}{S/\sqrt{n}}$ is the pivotal quantity around which the statistical inference regarding μ revolves, if the observations are made from a $n(\mu, \sigma^2)$ population with σ^2 unknown (see Chapters 11 and 12).

(b) Note that if v is large, the standard deviation of the t distribution is approximately equal to 1. In this case it can be proved that T approximately has a n(0, 1) distribution. However, this approximation holds good only if v is greater than 100.

9.3 The F distribution

As with the χ^2 distribution, the F distribution is positively skewed, while the F variable can also assume only positive values. This distribution is fully determined by two parameters, v_1 and v_2, that are known as the degrees of

freedom of the distribution. We also refer to the $F(\nu_1, \nu_2)$ distribution. The F distribution is of practical interest in the sampling distribution of the ratio of variances of samples from normal populations.

If two independent random samples of sizes n_1 and n_2 are drawn from a $n(\mu_1, \sigma_1^2)$ and a $n(\mu_2, \sigma_2^2)$ population, the sampling distribution of

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

is that of an $F(n_1 - 1, n_2 - 1)$ distribution, where S_1^2 and S_2^2 represent the sample variances of the two samples. The parameters $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ are called the numerator and denominator degrees of freedom.

From this result it follows, for instance, that if a sample of size 10 is drawn from a $n(\mu_1, 16)$ population and a sample of size 12 is drawn from a $n(\mu_2, 25)$ population, and the two sample variances are $s_1^2 = 14.7$ and $s_2^2 = 27.3$, then $f = \frac{14.7/27.3}{16/25} \approx 0.84$ is an observation of an F(9, 11) variable.

The density function of an F(12, 6) distribution is represented in Figure 10.17.

9.3.1 Tables of the F distribution

Suppose F_{ν_1, ν_2} indicates a stochastic variable that has the $F(\nu_1, \nu_2)$ distribution, then the customary F tables contain the values $F_{\nu_1, \nu_2; \alpha}$ for which

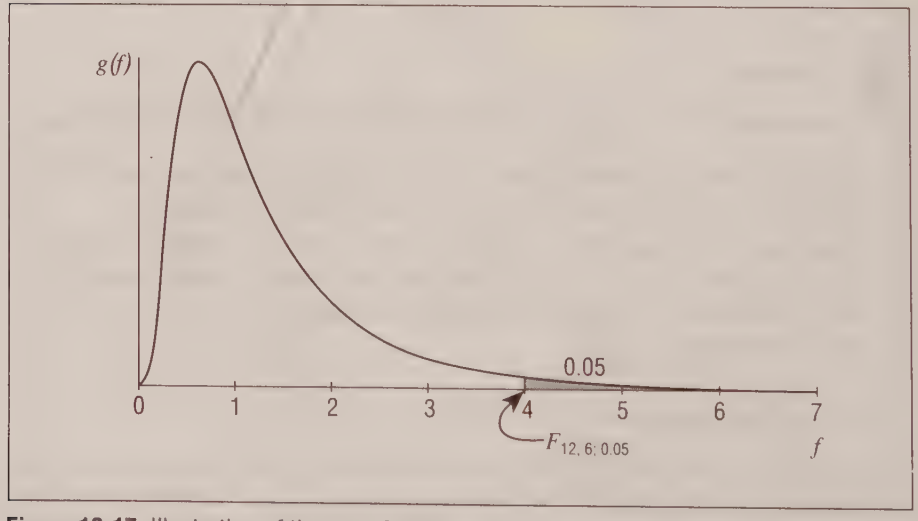


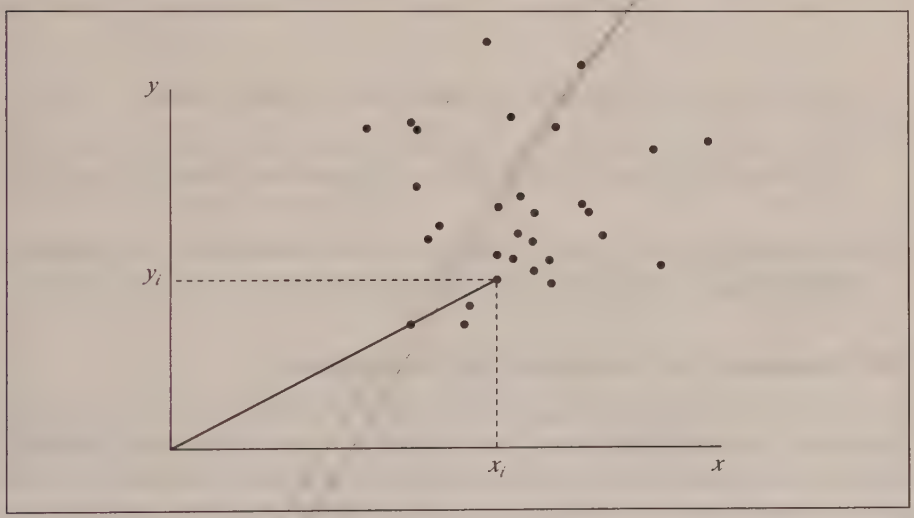
Figure 10.17 Illustration of the use of F tables

$$P\left(F_{\nu_1,\,\nu_2} \geq F_{\nu_1,\,\nu_2;\,\alpha}\right) = \alpha$$

for different values of ν_1, ν_2 and α. Table B.5 in Appendix B contains the $F_{\nu_1,\,\nu_2;\,\alpha}$ values for $\alpha \leq 0.10$. Figure 10.17 illustrates the interpretation of the F tables. For $\nu_1 = 12$, $\nu_2 = 6$ and $\alpha = 0.05$, it follows that $F_{12,\,6;\,0.05} = 4.00$.

EXAMPLE 10.25

An arms manufacturer tests the accuracy of a new hunting rifle by shooting at a target under constant conditions. The deviations in the horizontal (x) and vertical (y) directions should have the same degree of spread (see accompanying sketch). In a shooting test of sixteen shots it is found that $s_x^2 = 127\ (\text{mm})^2$ and $s_y^2 = 61\ (\text{mm})^2$. The manufacturer would like to determine the extent of the probability that $S_x^2 > 2S_y^2$ if the two population variances are equal, in other words if $\sigma_x^2 = \sigma_y^2$.



If we assume that the distribution of shots in both the x and y directions is normal and independent, it follows that

$$P\left(S_x^2 > 2S_y^2\right) = P\left(\frac{S_x^2}{S_y^2} > 2\right)$$

$$= P\left(\frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} > \frac{2\sigma_y^2}{\sigma_x^2}\right)$$

$$= P\left(F_{15,\,15} > 2\right), \quad \text{since } \sigma_x^2 = \sigma_y^2$$

From the F table it follows that

$P(F_{15,\,15} \geq 1.97) = 0.10$

$P(F_{15,\,15} \geq 2.40) = 0.05$

The required probability is therefore approximately 0.10. The manufacturer thus concludes that there is scant probability of obtaining a variance in the x direction that is twice the size of the variance in the y direction if the two population parameters σ_x^2 and σ_y^2 are equal.

Notes

(a) The quantity

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

is the pivotal quantity around which the statistical inference concerning the ratio $\frac{\sigma_1^2}{\sigma_2^2}$ revolves (see Chapters 11 and 12).

(b) The F tables are mainly used for values of $\alpha \leq 0.10$ and $\alpha \geq 0.90$. For values of $\alpha \geq 0.90$ the following relation can be used to determine $F_{\nu_1, \nu_2; \alpha}$.

$$F_{\nu_1, \nu_2; \alpha} = \frac{1}{F_{\nu_2, \nu_1; 1-\alpha}}$$

This relation follows because $\frac{1}{F_{\nu_1, \nu_2}}$ has an $F(\nu_2, \nu_1)$ distribution.

(c) If the F statistic has an $F(\nu_1, \nu_2)$ distribution, we can prove that $\nu_1 F$ approximately has a $\chi^2(\nu_1)$ distribution. This approximation is suitable for $\nu_2 \geq 100$.

(d) The F distribution also plays an important role in the analysis of variance. This important technique is discussed in Chapter 14.

10 Summary of sampling distributions

In the previous sections we discussed a variety of pivotal quantities and their associated sampling distributions, and pointed out that the statistical inference concerning the parameters μ, σ^2 and π are based on their distributions. For the sake of completeness we now give a brief summary of all the cases that are important to our purposes, including a number of cases that are now mentioned for the first time. We distinguish between two main groups of cases, namely the **one-sample problem** and the **two-sample problem**.

10.1 The one-sample problem

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a population with mean μ and variance σ^2.

10.1.1 Pivotal quantity for μ

The pivotal quantity $T = \frac{(\bar{X} - \mu)}{S/\sqrt{n}}$ has a $t(n-1)$ distribution provided the observations are taken from a normal population.

The pivotal quantity T approximately has a n(0, 1) distribution, provided n is large, regardless of the distribution of the population from which the sample was drawn.

10.1.2 Pivotal quantity for σ^2

The pivotal quantity $Y = \dfrac{(n-1)S^2}{\sigma^2}$ has a $\chi^2(n-1)$ distribution provided the observations come from a normal population.

10.1.3 Pivotal quantity for π

Suppose X is the number of successes in n mutually independent trials, with probability π of success, so that X has a bi(n, π) distribution, then

$$Z = \frac{P - \pi}{\sqrt{\dfrac{\pi(1-\pi)}{n}}}, \text{ with } P = \frac{X}{n},$$

will approximately have a n(0, 1) distribution, provided n is relatively large.

10.2 The two-sample problem

Suppose two independent random samples of sizes n_1 and n_2 are drawn from populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. The arithmetic means and variances of these two samples are indicated by $\bar{X}_1$, $\bar{X}_2$ and S_1^2, S_2^2.

10.2.1 Pivotal quantities for $\mu_1 - \mu_2$

The pivotal quantity

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

with $$S = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}}$$

has a t($n_1 + n_2 - 2$) distribution, provided the observations are from normal populations with **equal variances**.

The pivotal quantity

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$$

approximately has a n(0, 1) distribution, provided n_1 and n_2 are very large.

10.2.2 Pivotal quantity for $\frac{\sigma_1^2}{\sigma_2^2}$

The pivotal quantity

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

has an F($n_1 - 1$, $n_2 - 1$) distribution, provided the observations are from normal populations.

10.2.3 Pivotal quantity for $\pi_1 - \pi_2$

If $P_1 = \frac{X_1}{n_1}$ and $P_2 = \frac{X_2}{n_2}$ with X_1 and X_2 respectively, the number of successes in a bi(n_1, π_1) and bi(n_2, π_2) experiment, the pivotal quantity

$$Z = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{P_1(1 - P_1)}{n_1} + \frac{P_2(1 - P_2)}{n_2}}}$$

will approximately have a n(0, 1) distribution, provided n_1 and n_2 are both large.

Exercises

1 Refer to Exercise 9.2. Consider the random selection of one residence in Durban and define the stochastic variable X as follows:

$$X = \begin{cases} 1, \text{ if } A \text{ realizes (house on a small plot);} \\ 2, \text{ if } B \text{ realizes (house on a medium-sized plot);} \\ 3, \text{ if } C \text{ realizes (house on a large plot).} \end{cases}$$

Determine the probability distribution of X and present it in graphic form.

2 Repeat Exercise 1 for residences in Johannesburg and Pretoria by defining the respective stochastic variables Y and Z as in Exercise 1. Draw the two probability distributions on one set of axes.

3 (a) On the basis of the probability distribution of X in Exercise 1, determine:
   (i) the cumulative distribution function $F(b)$ for $1 \leq b \leq 3$;
   (ii) the mean and the variance of X, namely μ and σ^2.
   (b) Use (a) above to represent $F(b)$ in graphic form.

4 Repeat Exercise 3(a) and 3(b) for residences in Johannesburg and Pretoria (see also Exercise 2). Now draw the cumulative distribution functions of Y and Z, namely $F_Y(b)$ and $F_Z(b)$, on the same set of axes.

5 Refer to Exercise 9.3. Assume that the stochastic variable X is defined as follows:

$$X = \begin{cases} 2, \text{ if a red card is drawn;} \\ 4, \text{ if a yellow card is drawn; and} \\ 6, \text{ if a white card is drawn.} \end{cases}$$

Determine

(a) the probability distribution of X; and

(b) the cumulative distribution function $F(b)$ for $2 \leq b \leq 6$. Draw $F(b)$.

Also determine

(c) $P(X \geq 3)$ and $P(2 \leq X \leq 5)$; and

(d) μ and σ of X.

6 A stochastic variable Z has the following probability distribution:

$(1, \frac{1}{12}), (2, \frac{2}{12}), (3, \frac{3}{12}), (4, \frac{3}{12}), (5, \frac{2}{12}), (6, \frac{1}{12})$

(a) Present the distribution in graphic form and deduce from the graph the μ value for the mean of Z.

(b) Also determine

(i) $P(X \leq k)$ for $k = 1$ and 2;

(ii) $P(X \geq k)$ for $k = 5$ and 6. Explain the relation between these values and those of (i).

(c) Determine the values of μ and σ^2 by using the calculation formulas.

7 A discrete stochastic variable X has the following probability distribution:

$p(x) = \frac{k}{10}, \; x = 0, 1, \ldots, 9$

with k a constant value.

(a) Determine the value of k.

(b) If a stochastic variable Y has the following probability distribution, determine the value of the constant m:

$$p(y) = \begin{cases} m(y + 1), & y = 0, 1, \ldots, 4 \\ m(10 - y), & y = 5, 6, \ldots, 9 \end{cases}$$

8 A red and a black dice are rolled simultaneously. Let X = sum of the numbers yielded by the two dice.

Determine

(a) the probability distribution $p(x)$ of X and present it in graphic form;

(b) the cumulative distribution function $F(b)$ of X and present it in graphic form;

(c) the mean, μ, and the variance, σ^2, of X.

9 A discrete stochastic variable Y has the following cumulative distribution function:

$$F(b) = \begin{cases} 0.10, 0 \leq b < 1 \\ 0.25, 1 \leq b < 2 \\ 0.50, 2 \leq b < 3 \\ 0.75, 3 \leq b < 4 \\ 0.90, 4 \leq b < 5 \\ 1.00, 5 \leq b \end{cases}$$

Determine

(a) $P(Y = 1)$ and $P(Y \leq 2)$;

(b) $P(Y > 4)$;
(c) $P(2 < Y \leq 4)$; and
(d) μ and σ, the mean and standard deviation of Y.

10 A continuous stochastic variable X has the following probability model:

$$f(x) = \begin{cases} x, & 0 < x \leq 1 \\ 2 - x, & 1 < x < 2 \\ 0, & \text{otherwise} \end{cases}$$

(a) Draw the density curve for this distribution.
(b) By examining the density curve, determine the mean μ.
(c) Calculate $P(X < 0.25)$ and $P(X \geq 1.5)$.
(d) Determine the constant k for which $P(X \leq k) = 0.25$.

11 A continuous stochastic variable X has a density function which is given by

$$f(x) = \begin{cases} k - \frac{1}{8}x, & 0 < x < 8k \\ 0, & \text{otherwise} \end{cases}$$

with k a positive constant.
Determine
(a) the value of k;
(b) $F(x)$ for $x = 1$ and 3 (clue: first determine $P(X > x)$ for $x = 1$ and 3);
(c) $P(1 < X \leq x)$ from (b) above.

12 A continuous stochastic variable Y has the following cumulative distribution function

$$F(y) = \begin{cases} 0, & y < 0 \\ \frac{y^2}{64}, & 0 < y < 8 \\ 1, & y \geq 8 \end{cases}$$

Determine
(a) $P(Y \leq 4)$
(b) $P(Y > 6)$
(c) the value k for which $F(k) = 0.5$.

13 Assume that X has a uniform distribution over the interval (0, 10).
(a) Determine $F(b)$ for $b = 2, 4, 6, 8$ and 10. Plot the points $(b, F(b))$ for $b = 2, 4, \ldots, 10$. What do you notice?
(b) Use (a) above to find a general expression for $F(b)$.

14 Assume that X has a bi(n, π) distribution.
(a) If $n = 7$ and $\pi = 0.1$, determine $P(X = 4)$; $P(X \leq 1)$; μ and σ^2.
(b) If $n = 12$ and $\pi = 0.3$, determine $P(X = 2)$, $P(X \geq 4)$, $P(X \leq 6 \mid X \leq 8)$.
(c) If $n = 20$ and $\pi = 0.6$, determine $P(X = 10)$, $P(X \leq 8)$ and the greatest value b for which $F(b) \leq 0.10$.

15 Construct a bar chart of the bi(10, 0.2) distribution. Also determine the mean, μ, and the standard deviation σ.

16 From a container holding ten identical cards numbered 0, 1, …, 9, n cards are randomly drawn and replaced after noting the number. Let X = number of cards in the sample with number smaller than or equal to 3.
(a) Write down the probability function $p(x)$ and the domain of X if $n = 25$.
(b) Find $P(X \leq 10)$ and $P(2 \leq X \leq 8)$ for $n = 25$.
(c) Determine the minimum sample size n for which $P(X = 0) \leq 0.1$.

17 Fifty points are randomly selected from the rectangle between $x = -\frac{1}{2}$, $x = \frac{1}{2}$, $y = -\frac{1}{2}$ and $y = \frac{1}{2}$.
(a) Determine the probability distribution of X = number of points in the circle with centre 0 and radius $= \frac{1}{2}$.
(b) Determine the mean and standard deviation of X.
(c) Determine an approximate expression for $P(X \leq k)$.
(d) Find $P(30 \leq X \leq 45)$.

18 If X has a binomial distribution with mean $\mu = 7.2$ and variance $\sigma^2 = 2.88$, determine $P(X = 6)$.

19 An auditing firm knows from experience that 10% of the invoices from a certain retailer contain errors. In a random sample of 150 invoices 25 are found to contain errors.
(a) What is the probability that there will be errors in 25 or more invoices from this retailer?
(b) Determine the smallest value k for which the probability of finding at least k invoices with errors does not exceed 1%.

20 Soft drink containers are filled automatically and it is acceptable if 5% of the containers contain less than the quantity printed on the label. From a random sample of twenty containers, three are found to contain less than the required quantity. The supervisor maintains that this is 'normal' for a process of this nature. Do you agree?

21 The pass rate for a certain course is 65%. From a random sample of fifty students in this course, X is the number failing the course. Determine
(a) $P(X \geq 25)$;
(b) $P(20 < X < 30)$.

22 A multiple choice examination paper comprises 25 questions. Each question has four possible answers of which only one is correct. Suppose you answer the paper by randomly selecting an answer for each of the questions. Let Y be the number of questions that are correctly answered.
(a) What is the distribution of Y?
(b) Determine the probability
(i) that you will pass (at least 13 correct answers); or
(ii) that you will score zero.
(c) How much can you expect to score?

23 From a full pack of 52 playing cards, ten are drawn, noted and replaced. Let X = number of hearts among the ten cards drawn.
Determine
(a) the probability distribution of X, namely $p(x)$;
(b) μ and σ; and
(c) the probability that between four and eight cards (both included) will be hearts;
(d) $F(b)$ for $b = 0,\ 0.5,\ 1.0,\ 1.5,\ 2.0,\ 2.5,\ 3$.
(e) Present both $p(x)$ and $F(b)$ in graphic form.

24 Suppose X has a po(3) distribution. Determine
(a) $P(X = 0)$
(b) $P(X \leq 4)$
(c) $P(1 < X \leq 5)$

25 Suppose X has a po(λ) distribution with a probability function

$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!},\ x = 0,\ 1,\ 2,\ \ldots$$

Show that $p(x+1) = \frac{\lambda}{x+1}p(x),\ x = 0,\ 1,\ 2,\ \ldots$

26 On a certain road between two towns there are, on average, 4.5 potholes per kilometre. Determine the probability that there will be at least ten potholes over a distance of 2 km.

27 Suppose the probability that a person will suffer side effects after inoculation against a certain strain of influenza is 1 in 1 000. If 300 individuals are inoculated, determine the probability that
(a) no-one;
(b) exactly two people;
(c) at least two people will suffer side effects.

28 Determine approximate values of the cumulative distribution function $F(b)$ for $b \leq 10$ for the bi(30, 0.1) distribution by using Table B.7. Compare these values with the exact values derived from Table B.6.

29 A certain typist makes, on average, two typing errors per page. What is the probability that in a five-page document she will have
(a) at most five errors;
(b) exactly ten errors; and
(c) at least fifteen errors?

30 Assume that X has a hi(5, 10, 20) distribution. Determine
(a) $P(X = 3)$
(b) $P(X \leq 2)$
(c) μ and σ^2, the mean and the variance of X.

31 An auditor randomly selects the financial statements of five of the twelve most important clients of a particular enterprise in order to check the

particulars pertaining to their transactions. If there are errors in the transactions of four of the twelve clients, determine the probability that
(a) all four; and
(b) none
of these four clients' statements will be selected.

32 An auditing firm employs four ladies and eight men as clerks. Three of these clerks are selected to accompany an experienced auditor who is to audit the books of a client based at a coastal town. The three clerks selected are all men. The ladies maintain that they have been the victims of discrimination, but the manager maintains that he 'drew the names from a hat'. The ladies insist that such an outcome is 'improbable' and that the probability is less than 10%. Which argument is correct?

33 A student enrolling for a particular course has a probability P of passing the course the first time. Since P is not known, assume that this is a stochastic variable with a uniform distribution over the interval (0.40, 0.90).
(a) Give the mathematical function of the density curve $f(p)$.
(b) Determine the mean and the variance of P. What is the expected probability that he will pass?

34 Consider Chapter 5, Exercise 20, where the bookkeeping system of a sports club treasurer is discussed. If one can assume that the error in a transaction is uniformly distributed over the interval (–R0.50, R0.50) determine the probability of an error
(a) exceeding R0.45;
(b) lying between –R0.25 and R0.25; and
(c) exceeding R0.25, given that it is positive.

35 If X has an exponential distribution with mean $\theta = 5$, determine
(a) $P(X > 5)$ and $P(X > 5 \mid X > 4)$;
(b) $P(2 < X < 7)$; and
(c) the value k for which $P(X \leq k) = 0.5$. Interpret the difference between the values of θ and k.

36 On Saturday afternoons the demand for water from a reservoir has an exponential distribution with mean $\theta = 2.5$ cubic metres per second (cusec). Determine
(a) the probability that the demand for water will exceed 5 cusec on a specific Saturday afternoon; and
(b) the water delivery capacity required by the supply system to ensure that the probability of the capacity being exceeded by the demand is only 1%.

37 The life expectancy of a fan belt of a certain car make has an exponential distribution with mean $\theta = 60\,000$ km. What is the probability that the fan belt
(a) will last for longer than 80 000 km;
(b) will last for longer than 80 000 km if it has already done 60 000 km?

38 A fleet owner has ten vehicles of the make mentioned in Exercise 37. If these ten vehicles have already been in use for six months (with varying distances travelled), what is the probability that, during the next 10 000 km each of these vehicles will complete,
   (a) there will be no broken fan belts; or
   (b) at least two of the vehicles will have broken fan belts?

39 Vehicles drive past a specific point in one lane of a busy road at an average of 45 per minute.
   (a) Determine the mean time between two consecutive vehicles.
   (b) What is the probability of the time between two consecutive vehicles being
      (i) longer than 2 seconds; and
      (ii) between 1 and 2 seconds?

40 A bridge over a specific river must be designed for a fifty-year flood, i.e. the bridge must be designed to hold if the river reaches a level that is, on average, reached once in fifty years.
   (a) Assess the probability that the bridge will have to withstand a flood of this intensity within the next ten years.
   (b) Given that there has not been a fifty-year flood since the bridge was built fifteen years ago, what is the probability that a flood of this extent will occur during the coming year?

41 Suppose Y has a n(20, 36) distribution. Determine the following:
   (a) $P(Y \leq 15)$;
   (b) $P(25 < Y < 36)$;
   (c) $P(18 < Y < 30)$;
   (d) the value of k for which $P(Y < k) = 0.40$; and
   (e) the value of k for which $P(Y \geq k) = 0.90$.

42 The total amount X (in thousands of rand) claimed per annum by policy holders in a specific portfolio, has a n(1 850; 250 000) distribution. Determine
   (a) $P(X \leq 1\,000)$;
   (b) $P(X > 2\,500)$; and
   (c) the value r for which $P(X \geq r) = 0.01$.

43 Find the values of
   (a) $\Phi(1.5)$ and $\Phi(-1.5)$;
   (b) $\Phi^{-1}(0.75)$ and $\Phi^{-1}(0.1)$; and
   (c) $\Phi[\Phi^{-1}(0.5)]$

44 Suppose X has a n(10, 16) distribution. Determine
   (a) $P(5X > 20)$ and
   (b) $P(3X - 10 < 14)$

45 The weekly expenditure (in rand) on maintenance and repairs incurred by a photocopying and printing firm has a n(750, $(150)^2$) distribution. For one

specific week expenses amounted to R1 100. Is this unrealistically high? (Clue: Determine the probability of expenses exceeding R1 100 during any one week.)

46 The annual rainfall (in millimetre) in a certain area has a n(850, 10 000) distribution. The annual rainfall can be regarded as 'normal' if it deviates at most one standard deviation from the mean (to either side). Determine the 'normal' boundaries and the percentage time for which rain can be expected between these boundaries.

47 A machine that manufactures ball bearings must be set to produce bearings with a diameter that will lie within the specification limits of 2.000 ± 0.05 cm. At present the average of the bearings is 1.950 and the standard deviation 0.02.
(a) Which distribution would you assume for X, the diameter of the bearings? Motivate your answer.
(b) What percentage of the bearings does not meet the specification requirements at present?
(c) How much must the average diameter be enlarged or diminished to ensure a maximum percentage of bearings that will meet the specifications? How big is this percentage?

48 An employer with a large number of employees in his service knows from experience that the total number of hours lost per month as a result of labour unrest has a n(250; $(24)^2$) distribution.
(a) What is the probability that the time lost during the coming month will exceed 300 hours?
(b) How many lost hours should he plan for in his work schedules for the coming month if the probability of exceeding this number of planned hours is to be only 5%?

49 Refer to Exercise 34 and determine
(a) the mean and variance of $\bar{E}$, the average error in a random sample of 25 transactions (clue: $\mu_E = 0$, $\sigma^2_E = \frac{1}{12}$);
(b) $P(\bar{E} \leq -0.10)$
(c) $P(\bar{E} \geq 0.15)$.

50 If X has an exponential distribution with mean $\theta = 5$ and a random sample $X_1, X_2, \ldots, X_{100}$ is drawn from this distribution with $\bar{X} = \frac{\Sigma X}{100}$, determine
(a) $P(\bar{X} > 6.25)$
(b) $P(4 < \bar{X} < 7)$
(c) $P(\Sigma X < 300)$.

51 A specific brand of tyre gives an average of 40 000 km with a standard deviation of 5 000 km. A random sample of 60 tyres is tested. Calculate
(a) the mean and variance of $\bar{X}$, the average distance per tyre;
(b) $P(39\,000 < \bar{X} < 41\,000)$.

52 Consider Exercise 22. Suppose the examination paper comprises 100 questions. Determine
(a) the mean and the standard deviation of Y;
(b) $P(Y < 30)$;
(c) $P(Y = 20)$; and
(d) $P(15 < Y < 25)$.

53 Refer to Exercise 26 above and determine the probability that there will be more than 100 potholes over a distance of 25 km.

54 The probability that an individual will suffer a particular side-effect after inoculation against a certain strain of influenza is 0.001 (see Exercise 27). During an inoculation campaign, 15 000 people are inoculated. Determine the probability that more than 25 will actually suffer side-effects.

55 The following data represents the life expectancy (in years) of a specific electric component:

10.23, 3.92, 14.80, 0.26, 5.54, 0.27, 2.48, 0.76, 4.84, 3.76, 2.08, 5.17, 13.60, 2.41, 3.11, 13.12, 2.10, 1.71, 3.52, 1.98.

Draw a normal probability plot for the data and comment on whether the data is derived from a normal distribution.

56 The following data is presumed to have originated from a normal population. Does a normal probability plot confirm this presumption?

–2.38, –5.66, 3.87, –4.66, –1.29, 2.62, –2.85, –3.35, –4.81, –1.41, 1.96, 3.01, –4.45, –4.66, –1.52, –1.39, –2.58, –4.35, 1.12, 1.93.

57 Determine the following by using tables:
(a) $\chi^2_{24,\,0.05}$ and $\chi^2_{10,\,0.975}$
(b) a, for which $\chi^2_{15,\,\alpha}$ is as close as possible to 12
(c) $t_{12,\,0.05}$ and $-t_{25,\,0.025}$
(d) t, for which $P(|T_{10}| \geq t) = 0.20$
(e) ν, for which $t_{\nu,\,0.01} = 2.528$
(f) $F_{12,\,20;\,0.025}$ and $F_{12,\,20;\,0.975}$
(g) f, for which $P(F_{9,10} \geq f) = 0.05$
(h) $F_{8,\,10;\,0.90}$ and $F_{15,\,5;\,0.95}$

58 A random sample of size 10 is drawn from a $n(\mu_1, 2.25)$ distribution and an independent random sample of size 15 from a $n(\mu_2, 3.50)$ distribution. Let S_1^2 and S_2^2 respectively indicate the two sample variances. Determine
(a) $P(S_1^2 < 2.35)$
(b) $P(1.64 < S_2^2 < 5.92)$
(c) $P\left(\frac{S_1^2}{S_2^2} \geq 1.36\right)$
(d) a constant a so that $P\left(-a < \frac{\overline{X} - \mu_1}{S} < a\right) = 0.90$

CHAPTER 11

Estimation

1 Introduction

Estimation is part of our everyday lives. The housewife buying her monthly groceries makes an estimate of how much of each product she has to buy. If she wants to pay cash for these purchases, she has to **estimate** the cost beforehand to make sure she does not have too little money. The treasurer of an association has to **estimate** the association's expenditure over the next few months before he can invest R10 000 of the association's funds. The manager of a construction company makes an **estimate** of the expected company profits in the coming year before he decides on salary adjustments, purchasing new heavy vehicles and employing more labourers. A pharmaceutical company makes an **estimate** of the percentage of clients that will benefit from a new drug before deciding whether or not to market the drug for the disease concerned.

It is therefore clear that any decision taken by a person or an institution is usually preceded by an estimate. To a large extent, the correctness of the decision is determined by the **accuracy** of the estimate. However, an estimate of an unknown quantity is impossible if no relevant information is available. Thus, for example, the manager of the construction company bases his estimate of future profit on the profits made over the past few years, with due regard for the economic climate (see Chapter 7). The pharmaceutical firm, on the other hand, tests the new drug on a sample of patients. The proportion of patients who benefit then leads to an estimate of the effectiveness of this drug.

In this chapter only the estimate of a population mean, a population proportion and a population variance will be discussed. We shall assume throughout that the relevant information on which the estimates of these **population parameters** are based is obtained from a simple random sample of elements from the population concerned. There are two ways of estimating a population parameter:

A **point estimate** is a single numerical value that is used to estimate an unknown population parameter.

An **interval estimate** is a set of consecutive values that is used to estimate an unknown population parameter with a specified degree of certainty.

Although a point estimate can provide important information on a population parameter, it almost always differs from the unknown parameter. The population parameter can be estimated correctly only if the population as a whole is investigated. An interval estimate, on the other hand, indicates a set of values in the neighbourhood of the point estimate that will contain the unknown parameter with a certain probability.

Suppose the mean floor area (m^2) of residences in the Johannesburg area is considered in an investigation. As the 843 newly purchased residences in the area (see Appendix A, Data Set 1) are representative of all houses in Johannesburg with regard to this specific population characteristic, a simple random sample from this population can provide the necessary information. Such a sample of size 30 provides $\bar{x} = 175.30\ m^2$, and $s = 74.80\ m^2$. The arithmetic mean $\bar{x} = 175.30\ m^2$ is now a point estimate of the mean area μ. By means of calculations (148.50, 202.10) is found to be an interval estimate for μ. Although this is a very wide interval of possible values of μ, the probability associated with this is 0.95. This is known as a 95% confidence interval for μ (see Section 3). Since the information contained in the sample is used to shed light on a population parameter, this is a case of statistical inference, or statistical generalization. Point and interval estimates form an integral part of statistical inference. These two techniques will now be discussed fully.

2 Point estimation

2.1 The population mean μ

Suppose X is a stochastic variable associated with a feature or characteristic of the population and that μ, the mean of the distribution of X, is unknown. Let $X_1, X_2, \ldots, X_n$ represent a simple random sample from this population.

The **customary point estimator of μ** is given by

$$\bar{X} = \frac{\sum X}{n}$$

The word 'estimator' indicates the stochastic variable $\bar{X}$, which is a function of the stochastic variables $X_1, X_2, \ldots, X_n$. On the other hand, the 'estimate' $\bar{x} = \frac{\sum x}{n}$ can be calculated from the observed sample $x_1, x_2, \ldots, x_n$.

2.2 The population variance σ^2

Suppose the variance of the distribution of X is unknown and that $X_1, X_2, \ldots, X_n$ is a simple random sample from the population concerned.

The **customary point estimator of σ^2** is given by

$$S^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

2.3 The population proportion π

Let π be the unknown proportion of persons or objects in a population with a certain characteristic. A simple random sample of n entities from this population gives the number X with this specific characteristic.

The **customary point estimator of π** is given by

$$P = \frac{X}{n}$$

EXAMPLE 11.1

Regard the 376 values of the floor areas of residences in Durban as a random sample of the population of areas of all residences in this city (see Appendix A, Data Set 1). Calculations yield

$$\sum x = 65\,400, \quad \sum x^2 = 12\,752\,100.$$

From this it follows that

$$\bar{x} = \frac{65\,400}{376} = 173.94 \text{ m}^2$$

and

$$s^2 = \frac{1}{375}\left(12\,752\,100 - \frac{(65\,400)^2}{376}\right)$$

$$= 3\,671.13 \left(\text{m}^2\right)^2$$

These values are now the respective point estimates of the unknown mean μ and the variance σ^2 of floor areas of residences in Durban. If an estimate of the proportion of houses with an area that exceeds 200 m^2 is also required, the relative frequency of the number of houses among the 376 with this characteristic can be used as a point estimate, namely

$$p = \frac{118}{376} = 0.314$$

2.4 Characteristics of point estimators

The point estimators proposed above for the unknown parameters μ, σ^2 and π are intuitively acceptable in view of the subject matter discussed in previous chapters. The question, however, is: are there other point estimators for these parameters that provide just as much or even more information on the unknown parameters? Is the median not perhaps a better estimator of μ if the sample is drawn from a symmetrical distribution? How does $M_2 = \frac{\sum (X - \bar{X})^2}{n}$ compare with $S^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$ as an estimator of σ^2? To answer such questions, objective criteria have to be found to evaluate a point estimator. We shall discuss three criteria that are essential to a good estimator.

2.4.1 Unbiasedness

An estimator of an unknown population parameter is known as an **unbiased estimator** if the mean of the sample distribution of the estimator is equal to the population parameter concerned.

If samples of the same size are repeatedly drawn from a population and an estimate of the unknown population parameter is made in each case, the average value of these estimates will be equal to the population parameter. From Chapter 10, Section 8.1, it follows that $\bar{X}$ is an unbiased estimator of μ. It can also be shown that S^2 and P are respectively unbiased estimators of σ^2 and π.

2.4.2 Efficiency

The standard deviation of the sample distribution of an estimator is just as important as the mean of this distribution. If we had to choose between two unbiased estimators of an unknown population parameter, we would choose the one that has a sample distribution with the smallest standard deviation for a given value of n. The unbiased estimator that fluctuates least from one sample to the next is naturally the most reliable.

An unbiased estimator of an unknown population parameter is called an **efficient estimator** of the parameter concerned if it has the smallest standard deviation of all unbiased estimators for a given value of n.

We already know that the standard deviation of the sample distribution of $\bar{X}$ is equal to $\frac{\sigma}{\sqrt{n}}$. As we can show that this value is usually smaller than the standard deviations of all other unbiased estimators of μ, it follows that $\bar{X}$ is usually an efficient estimator of μ (however, see also Chapter 16, Section 7, Note (c)). We can also show that P is an efficient estimator of π and that the variance of S^2 for large samples does not differ much from that of the efficient estimator.

2.4.3 Consistency

A point estimator is called a **consistent estimator** if the difference between this estimator and the unknown population parameter with a probability of almost 1 can be made smaller than a preselected value by sufficiently enlarging the sample.

If an estimator is consistent, it is important to draw as large a sample as possible to obtain maximum accuracy.

However it is a waste of time and money to enlarge the sample if a non-consistent estimator is to be used to estimate the unknown population parameter. The estimators of μ, σ^2 and π, namely, $\bar{X}$, S^2 and P respectively are all consistent estimators.

Notes

(a) As S^2 and M_2 are virtually equivalent for large samples, it follows that M_2 is also a consistent estimator of σ^2.

(b) Since the standard deviation of an estimator is a function of the population parameters it is practically unknown. However, this important quantity can also be estimated by means of an observed sample and is then known as the **standard error of estimation**. Thus, for example, $\frac{s}{\sqrt{n}}$ is the standard error of estimation of $\bar{X}$ as the estimator of μ, while the standard error of estimation of P as the estimator of π is equal to $\sqrt{\frac{p(1-p)}{n}}$. For an observed sample of $x_1, x_2, \ldots, x_n$ both the point estimate $\bar{x}$ of μ and the standard error of estimation $\frac{s}{\sqrt{n}}$ can be determined.

(c) The **bias** of an estimator is defined as the difference between the mean of the sample distribution of the estimator and the particular population parameter.

(d) When an estimator is assessed, all its characteristics should be studied. An estimator can, for example, be biased, but have a smaller standard deviation than other estimators that are unbiased. There are, however, statistical criteria that indicate the preferred estimator in such cases. The best-known of these is the **mean squared error** (*mse*) of an estimator, which is defined as:

$$mse = (\text{bias})^2 + (\text{standard deviation of estimator})^2$$

EXAMPLE 11.2

Twenty-five simple random samples of size 10 are drawn from the 'population' of 843 residences in Johannesburg and the prices per square metre determined. The sample means and medians are given in Table 11.1.

Table 11.1 Means and medians of 25 samples of size 10 each from the population of 843 house prices (R/m²) (see Appendix A, Data Set 1)

Sample number	$\bar{x}$	me	Sample number	$\bar{x}$	me
1	915	912	14	979	1 001
2	833	795	15	940	878
3	813	776	16	911	919
4	753	739	17	933	809
5	884	909	18	982	940
6	1 041	975	19	747	715
7	1 144	1 183	20	903	820
8	951	880	21	785	795
9	884	924	22	906	905
10	974	849	23	944	924
11	781	800	24	779	749
12	1 013	1 088	25	1 141	1 199
13	874	1 002			

From the data contained in the table it follows that

$$\sum \bar{x} = 22\,810; \quad \sum \bar{x}^2 = 21\,084\,616$$

$$\sum me = 22\,486; \quad \sum (me)^2 = 20\,607\,766$$

Estimates of the means of the sample distributions of $\bar{X}$ and the median are given respectively by

$$\bar{x}_{\bar{x}} = \text{R}913/\text{m}^2, \; \bar{x}_{me} = \text{R}899/\text{m}^2$$

The estimated standard deviations of the sample distributions of $\bar{X}$ and the median as estimators of the unknown population parameter μ are:

$$s_{\bar{x}} = \text{R}107/\text{m}^2 \text{ and } s_{me} = \text{R}126/\text{m}^2$$

The true mean of the population price per square metre of the 843 houses is R922/m². It seems that $\bar{x}_{\bar{x}}$ does not differ much from μ and this is in agreement with a well-known statistical result, namely that $\bar{X}$ is always an unbiased estimator of μ. As the number of samples (all of size 10) increases, $\bar{x}_{\bar{x}}$ and μ will differ even less. The fact that $\bar{x}_{me}$ is considerably smaller than μ indicates that the median is a biased estimator of μ. This bias indicates that the distribution of house prices per m² for the Johannesburg area could be positively skewed (see also Chapter 5, Figure 5.6). Apart from the fact that the median is a biased estimator of μ in this case, the estimated standard deviation of the median is also greater than that of $\bar{X}$. From note (d) above it follows that the preferred estimator for this situation is the one that has the smallest mean squared error. For $\bar{X}$ as the estimator of μ it follows that the estimated mean squared error is given by

$$mse(\bar{x}) = (913 - 922)^2 + (107)^2$$
$$= 11\,530$$

For the median as the estimator of μ it follows that the mean square error is given by

$$mse(me) = (899 - 922)^2 + (126)^2$$
$$= 16\,405$$

Since $mse(\bar{x}) < mse(me)$, this criterion also gives preference to the arithmetic mean over the median as the estimator of μ for this particular population.

3 Interval estimation

A point estimator with a substantial standard deviation can yield an estimate that is too far off the mark. On the other hand, a point estimator with a small standard deviation is very accurate, although it is highly improbable that it will lead to the exact estimation of the unknown parameter. In view of this argument, an interval estimate rather than a point estimate of the population parameter is preferable. We shall now derive interval estimates for μ, σ^2 and π.

3.1 Confidence interval for μ

Suppose $X_1, X_2, \ldots, X_n$ is a simple random sample from a $n(\mu, \sigma^2)$ population, with both parameters unknown. From Chapter 10, Section 10 it follows that the pivotal quantity $T = \dfrac{\bar{X} - \mu}{S/\sqrt{n}}$, with $\bar{X}$ and S based on the sample $X_1, X_2, \ldots, X_n$, has a $t(n-1)$ distribution. Consequently,

$$P\left(-t_{n-1,\,0.025} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{n-1,\,0.025}\right) = 0.95$$

When calculations for inequalities are used, it follows that

$$P\left(\bar{X} - t_{n-1,\,0.025}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,\,0.025}\frac{S}{\sqrt{n}}\right) = 0.95$$

If $x_1, x_2, \ldots, x_n$ is an observed random sample from a $n(\mu, \sigma^2)$ population, a 95% **confidence interval for** μ is given by

$$\left(\bar{x} - t_{n-1,\,0.025}\frac{s}{\sqrt{n}},\ \bar{x} + t_{n-1,\,0.025}\frac{s}{\sqrt{n}}\right)$$

If the above confidence interval is calculated for observed random samples of size n from a $n(\mu, \sigma^2)$ population, on average 95 of each 100 such intervals will contain the unknown population parameter μ. The value 95% is known as the **confidence coefficient** of the interval.

EXAMPLE 11.3

For each of the 25 observed random samples of size 10 from the population of 843 house prices (see Example 11.2) a 95% confidence interval for the mean μ is calculated by means of

$$\left(\bar{x} - t_{9,\,0.025}\frac{s}{\sqrt{n}},\ \bar{x} + t_{9,\,0.025}\frac{s}{\sqrt{n}}\right)$$

where $t_{9,\,0.025} = 2.262$ (see Table B.3).

For the first sample

$\bar{x} = 915$ and $s = 313$.

The 95% confidence interval for μ is therefore:

$$\left(915 - 2.262\frac{313}{\sqrt{10}},\ 915 + 2.262\frac{313}{\sqrt{10}}\right)$$

i.e. (915 − 223.9, 915 + 223.9)
i.e. (691, 1 139).

Table 11.2 shows this confidence interval and the other 24 confidence intervals calculated similarly from the other samples.

Table 11.2 95% confidence intervals for μ calculated for each of the 25 observed random samples of size 10 (see Example 11.2)

Sample number	Confidence interval	Sample number	Confidence interval
1	(691, 1 139)	14	(737, 1 221)
2	(651, 1 015)	15	(728, 1 152)
3	(618, 1 008)	16	(695, 1 127)
4	(506, 1 000)	17	(660, 1 206)
5	(569, 1 199)	18	(778, 1 186)
6	(657, 1 425)	19	(521, 973)
7	(770, 1 518)	20	(628, 1 178)
8	(707, 1 195)	21	(651, 919)
9	(621, 1 147)	22	(699, 1 113)
10	(702, 1 246)	23	(744, 1 144)
11	(584, 978)	24	(601, 957)
12	(831, 1 195)	25	(901, 1 381)
13	(568, 1 180)		

We agreed tacitly that the stochastic variable X, namely house price per square metre, has a $n(\mu, \sigma^2)$ distribution. Without this assumption the above confidence interval for μ cannot be calculated from samples of size 10. The bias of the median as the estimator of μ (see Example 11.2) suggests that this assumption could be incorrect. However, it is known that smaller deviations from normality will not have much effect on the confidence coefficient obtained with the aid of the t distribution. The confidence intervals in Table 11.2 are represented graphically in

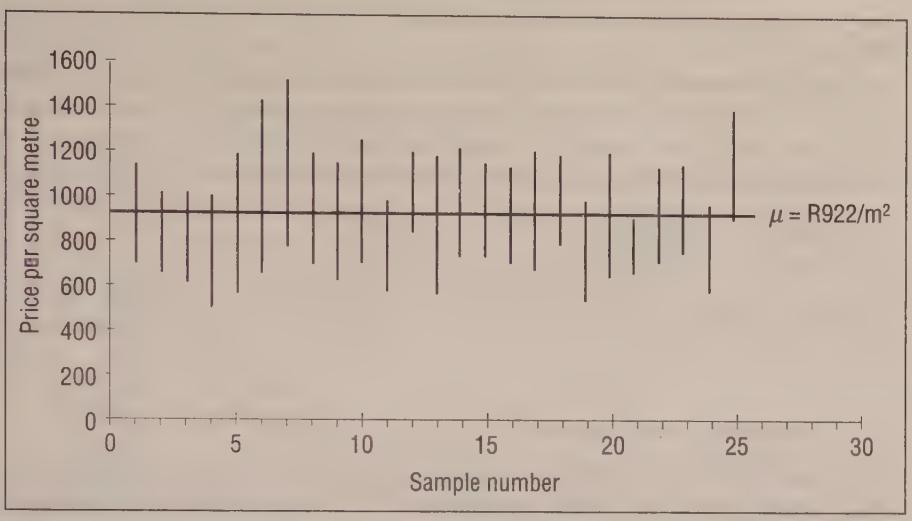


Figure 11.1 Graphic representation of 95% confidence intervals for the population mean μ (see Table 11.2)

Figure 11.1. Table 11.2 indicates that all the intervals except one contain the value $\mu = 922$ (see also Figure 11.1). This is in agreement with the theory according to which approximately 95 of every 100 such intervals will contain μ.

3.2 Confidence interval for σ^2

Suppose that $X_1, X_2, \ldots, X_n$ is a simple random sample from a $n(\mu, \sigma^2)$ population, in which the population variance σ^2 is unknown. From Chapter 10, Section 10 it follows that the pivotal quantity $Y = \frac{(n-1)S^2}{\sigma^2}$ has a $\chi^2(n-1)$ distribution. From this it follows that

$$P\left(\chi^2_{n-1,\,0.975} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1,\,0.025}\right) = 0.95$$

Simple inequality operations give

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1,\,0.025}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,\,0.975}}\right) = 0.95$$

If $x_1, x_2, \ldots, x_n$ is an observed random sample from a $n(\mu, \sigma^2)$ population, a **95% confidence interval for σ^2** is given by

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1,\,0.025}}, \frac{(n-1)s^2}{\chi^2_{n-1,\,0.975}}\right)$$

EXAMPLE 11.4

For the first observed sample of size 10 used in Examples 11.2 and 11.3 we find that $\bar{x} = 915$ and $s^2 = 98\,049$. According to Table B.4 $\chi^2_{9,\,0.975} = 2.700$ and $\chi^2_{9,\,0.025} = 19.023$. If we therefore assume that this is an observed sample from a normal population, it follows that a 95% confidence interval for σ^2 is given by

$$\left(\frac{9(98\,049)}{19.023}, \frac{9(98\,049)}{2.700} \right)$$

i.e. (46 388, 326 830).

3.3 Confidence interval for π

If π is the unknown proportion of persons or entities in a population with a certain characteristic, the sample proportion $P = \frac{X}{n}$ can be used as an estimator of π. The stochastic variable X is the number of elements in a sample of n entities with the specific characteristic. From Chapter 10, Section 10 it follows that the pivotal quantity

$$Z = \frac{P - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

approximately has a n(0, 1) distribution, provided n is reasonably large. Using Table B.2 it therefore follows that

$$P\left(-1.96 \leq \frac{P - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \leq 1.96 \right) \approx 0.95$$

If, in the expression $\sqrt{\frac{\pi(1-\pi)}{n}}$, π is replaced by its point estimator P, it follows (after further manipulation) that the following expression holds approximately:

$$P\left(P - 1.96\sqrt{\frac{P(1-P)}{n}} \leq \pi \leq P + 1.96\sqrt{\frac{P(1-P)}{n}} \right) \approx 0.95$$

If $p = \frac{x}{n}$ is the observed sample proportion and n is relatively large, an approximate **95% confidence interval for** π is given by

$$\left(p - 1.96\sqrt{\frac{p(1-p)}{n}}, \; p + 1.96\sqrt{\frac{p(1-p)}{n}} \right)$$

In situations where the value of π is thought to be close to 0 or 1, and where n is small, the normal approximation of P will not be good. Under these circumstances the confidence coefficient of the above interval can differ considerably from 95%.

Note

The above 95% confidence interval for π follows because, in the two relevant inequalities, the expression $\sqrt{\frac{\pi(1-\pi)}{n}}$ is replaced by $\sqrt{\frac{P(1-P)}{n}}$ for the sake of simplicity. However, we find that these two inequalities lead to the following quadratic inequality:

$$\frac{(P-\pi)^2}{\frac{\pi(1-\pi)}{n}} \leq (1.96)^2$$

Solving this in terms of π gives the following 95% confidence interval

(P_L, P_U) with

$$P_L = \frac{(1.96)^2 + 2np - \sqrt{(1.96)^4 + 4(1.96)^2 np(1-p)}}{2\left(n + (1.96)^2\right)}$$

$$P_U = \frac{(1.96)^2 + 2np + \sqrt{(1.96)^4 + 4(1.96)^2 np(1-p)}}{2\left(n + (1.96)^2\right)}$$

For small values of the sample size n this approximation provides a more accurate answer.

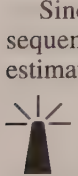
EXAMPLE 11.5

In Example 11.1 we see that $p = \frac{118}{376} = 0.314$ is the observed sample proportion of 376 residences in Durban with a floor area of more than 200 m^2. An approximate 95% confidence interval for π, the unknown proportion of all residences in Durban with this characteristic, is therefore

$$\left(0.314 - 1.96\sqrt{\frac{0.314(1-0.314)}{376}},\ 0.314 + 1.96\sqrt{\frac{0.314(1-0.314)}{376}}\right)$$

i.e. (0.27, 0.36).

Since $n = 376$, the normal approximation of P will be satisfactory and consequently 95% will be a true reflection of the probability related to this interval estimator of π.



General notes on interval estimation

(a) Although interval estimation with a confidence coefficient of 95% is commonly used, 99% or 90% confidence intervals can be similarly obtained.

Thus, for example, a 90% confidence interval for μ is given by

$$\left(\bar{x} - t_{n-1,\,0.05}\frac{s}{\sqrt{n}},\ \bar{x} + t_{n-1,\,0.05}\frac{s}{\sqrt{n}}\right)$$

while a 99% confidence interval for μ is given by

$$\left(\bar{x} - t_{n-1,\,0.005}\frac{s}{\sqrt{n}},\ \bar{x} + t_{n-1,\,0.005}\frac{s}{\sqrt{n}}\right)$$

An approximate 90% confidence interval for π is

$$\left(p - 1.645\sqrt{\frac{p(1-p)}{n}},\ p + 1.645\sqrt{\frac{p(1-p)}{n}}\right)$$

It is important to note that for a given sample size the length of the confidence interval increases as the confidence coefficient increases. The following are respectively 90%, 95% and 99% confidence intervals for μ with respect to the first sample of size 10 used in Example 11.2:

Confidence coefficient	Confidence interval	Interval length
90%	(734, 1 096)	362
95%	(691, 1 139)	448
99%	(593, 1 237)	644

Since a confidence coefficient of 90% is unacceptable to most people, and the length of a 99% confidence interval is usually too large to be of practical use, a 95% confidence interval is usually the selected compromise.

(b) All three the interval estimates discussed in this section are known as central confidence intervals. Suppose an interval estimate (a, b) does not contain the population parameter μ. In a central confidence interval there is an equal probability that μ will be greater than b or smaller than a. In some practical situations the consequences of a parameter value that is greater than the upper limit of the estimated interval, are far more serious than for a parameter value that is smaller than the lower limit. In such cases **non-central confidence intervals** are often used. The so-called **lower-tail** and **upper-tail** (or **one-sided lower** and **one-sided upper**) **confidence intervals** are included in this category of confidence intervals.

(c) Due to the symmetry of the underlying probability distributions the point estimates $\bar{x}$ and p are the midpoints of the confidence intervals for μ and π. The lengths of these two 95% confidence intervals are therefore:

$$\mu\text{:}\quad 2 \times t_{n-1,\,0.025}\,\frac{s}{\sqrt{n}}$$

$$\pi\text{:}\quad 2 \times 1.96\sqrt{\frac{p(1-p)}{n}}$$

(d) Since the lengths of the above two intervals are inversely proportional to $\sqrt{n}$, and since $t_{n-1,\,0.025}$ decreases for every increase in n (see Table B.3), it follows that the confidence intervals for μ and π become increasingly shorter as larger samples are used in the estimation. The shorter the interval estimates, the more information they provide on the unknown population parameter. Although not immediately apparent, the confidence interval for σ^2 has the same characteristic.

(e) In the derivation of the confidence intervals for μ and σ^2 we assumed that the sample was from a $n(\mu, \sigma^2)$ population. If this is not the case, the confidence coefficients of the intervals obtained can be queried. However, from the central limit theorem it follows that the confidence coefficient of the 95% confidence interval for μ for large samples will not differ much from 95%. Since the t distribution for large samples is well approximated by the n(0, 1) distribution, we can use the following very general result:

> If $x_1, x_2, \ldots, x_n$ is an observed random sample from a population with unknown mean μ, a **95% confidence interval for** μ for large samples is given by
>
> $$\left(\bar{x} - 1.96\,\frac{s}{\sqrt{n}},\ \bar{x} + 1.96\,\frac{s}{\sqrt{n}}\right)$$

4 Determining sample size

Throughout the discussion of point and interval estimation we assumed that the sample size is a fixed given value. However, when planning a research project, the researcher is usually faced with the question: "How large should the sample be?" As already stated, the confidence intervals for μ, π and σ^2 become shorter if n is increased.

In many research projects it is too expensive or time-consuming to draw a large sample. In quality control, items are often destroyed when they are tested. It can also, for example, take many years to involve a large number of patients with a specific disease in medical research. These factors often mean that researchers merely 'close their eyes and hope for the best' and believe that a relatively small sample will suffice.

However, it is important to remember that an inaccurate estimate of a population parameter can cost the researcher dearly. A small quality control sample can give the producer the impression that the items produced are of a high quality. Imagine his dismay when retailers return one consignment after

another because of poor quality. Since the cost involved in a sample of a certain size can seldom be precisely predetermined, it is better to ask: "What is the minimum sample size for estimating the parameter that will have a prescribed accuracy with a prescribed confidence coefficient?" rather than the usual "How much will it cost?" We now answer the former question for cases in which the unknown population mean and the population proportion have to be estimated.

4.1 Determining sample size when estimating μ

Suppose a building society branch needs to estimate the mean house valuation in Johannesburg in the fourth quarter of 1991. Suppose also that the building society concerned has access to the information on the 843 valuations contained in Appendix A, Data Set 1, and that they want to draw a sample from this data. Such an estimate provides valuable information for the evaluation of the loan amounts granted by the branch. The branch manager stipulates that the sample average house valuation with a probability of 0.95 may differ with only R20 000 from the real mean μ. From this it follows, with $\bar{X}$ as the point estimator, that

$$P(|\bar{X} - \mu| \leq 20\,000) = 0.95$$

or

$$P(-20\,000 \leq (\bar{X} - \mu) \leq 20\,000) = 0.95$$

Suppose σ is the population standard deviation of the valuations, then from the above equation it follows that

$$P\left(\frac{-20\,000}{\frac{\sigma}{\sqrt{n}}} \leq \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}} \leq \frac{20\,000}{\frac{\sigma}{\sqrt{n}}}\right) = 0.95$$

If $Z = \dfrac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$, then

$$P\left(\frac{-20\,000}{\frac{\sigma}{\sqrt{n}}} \leq Z \leq \frac{20\,000}{\frac{\sigma}{\sqrt{n}}}\right) = 0.95$$

From the central limit theorem it follows that, for large samples, Z has approximately a n(0, 1) distribution. Using Table B.2 it therefore follows that

$$\frac{20\,000}{\frac{\sigma}{\sqrt{n}}} = 1.96$$

Therefore $n = \left(\dfrac{1.96\,\sigma}{20\,000}\right)^2$

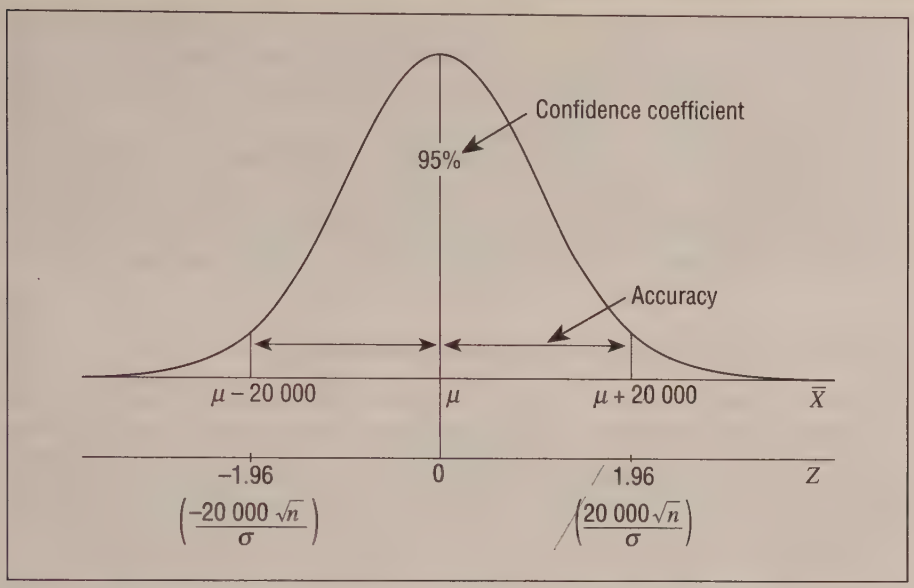


Figure 11.2 Relation between sample size, confidence coefficient and accuracy in estimating μ

Figure 11.2 represents the relation between the confidence coefficient, the accuracy and the sample size. However, the value of n cannot be determined, since σ is still unknown. A sensible plan would be to draw a simple random sample of size 30 valuations, to estimate σ by means of the sample standard deviation s, and then to calculate n. If n is greater than 30, an additional $(n - 30)$ house valuations are drawn randomly from the remaining population valuations. For an initial sample of 30 house valuations the following data was obtained: $\bar{x}$ = R161 400, s = R88 800. From this it is approximated that

$$n = \left(\frac{1.96 \times 88\,800}{20\,000}\right)^2$$
$$\approx 76$$

It is therefore essential to select another 46 valuations from the remaining 804 valuations. The final sample of size 76 gives

$$\bar{x} = \text{R}167\,400$$

We are now 95% confident that, during the fourth quarter of 1991, the real mean valuation of residences in Johannesburg did not differ from R167 400 by more than R20 000. This means that if this process were to be carried out many times, the difference between μ and $\bar{X}$ would not be more than R20 000 in 95% of the cases.

The following general result for determining the sample size therefore applies:

If $\bar{X}$ is used to estimate μ and the difference between $\bar{X}$ and μ is required to be no greater than e with a probability of 0.95, then by approximation $n = \left(\frac{1.96\,\sigma_b}{e}\right)^2$, with σ_b the population standard deviation σ (if σ is known), or an estimated value thereof (if σ is not known).

4.2 Determining sample size when estimating π

A pharmaceutical company wants to determine accurately, to within 2%, the percentage of patients who react positively to treatment with a new drug for a certain disease. The company is prepared to run a 5% risk that the estimated percentage will differ by more than 2% from the correct value, namely $100\pi\%$. The proportion P of patients (from a sample of n) who react positively to the treatment is used as an estimator of π. The following is therefore required:

$$P(|P - \pi| \leq 0.02) = 0.95$$

or

$$P(-0.02 \leq (P - \pi) \leq 0.02) = 0.95$$

Through standardization we obtain

$$P\left(\frac{-0.02}{\sqrt{\frac{\pi(1-\pi)}{n}}} \leq \frac{(P-\pi)}{\sqrt{\frac{\pi(1-\pi)}{n}}} \leq \frac{0.02}{\sqrt{\frac{\pi(1-\pi)}{n}}}\right) = 0.95$$

or

$$P\left(\frac{-0.02}{\sqrt{\frac{\pi(1-\pi)}{n}}} \leq Z \leq \frac{0.02}{\sqrt{\frac{\pi(1-\pi)}{n}}}\right) = 0.95$$

If the normal approximation for the distribution of P is satisfactory, it follows that Z is approximately distributed as n(0, 1). From Table B.2 it therefore follows that

$$\frac{0.02}{\sqrt{\frac{\pi(1-\pi)}{n}}} = 1.96$$

or

$$n = \pi(1-\pi)\left(\frac{1.96}{0.02}\right)^2$$

As the sample size n is expressed in terms of the unknown parameter π, it cannot be determined out of hand. However, the pharmaceutical company can follow one of two methods:

(a) Apply initial treatment to a number of patients, say fifty, with the new medicine, estimate π by means of the observed sample proportion p, and determine n from the above formula. If n is greater than 50, a further $(n - 50)$ patients are treated and a final estimate of π is made on the basis of the sample proportion of patients who reacted positively in the total sample.

(b) As $\pi(1 - \pi) \leq \frac{1}{4}$ for all values of π, it follows that

$$\frac{1}{4}\left(\frac{1.96}{0.02}\right)^2 = 2\,401$$

gives an upper bound to the desired sample size. If, therefore, 2 401 patients are involved in the investigation, the pharmaceutical company can be sure that the accuracy of the sample ratio does meet the requirements. This procedure is independent of the actual value of π, but often leads to needlessly large samples.

Suppose, for example, that π is actually 0.85, then, strictly speaking, a sample of the following size is needed:

$$n = 0.85 \times 0.15\left(\frac{1.96}{0.02}\right)^2 \approx 1\,225$$

In summary, the following general result applies:

If P is used to estimate π and the difference between P and π is required to be no greater than e with a probability of 0.95, then, by approximation, $n = \pi_b(1 - \pi_b)\left(\frac{1.96}{e}\right)^2$ applies, with π_b as a provisional estimate of π, provided the distribution of P can be satisfactorily approximated by the normal distribution.

Notes

(a) Obviously, if a confidence coefficient of 99% rather than one of 95% is required for the prescribed accuracy of the point estimators $\bar{X}$ and P, the two formulas for n will be given respectively by

$$n = \left(\frac{2.58\,\sigma_b}{e}\right)^2 \text{ and } n = \pi_b(1 - \pi_b)\left(\frac{2.58}{e}\right)^2$$

(b) Where small samples are drawn from a normal population with a view to estimating μ, there is no explicit formula for determining n. The same applies to the determination of n for estimating σ^2 for the normal population. However, in these two cases the sample sizes can be found by iteration.

(c) It is sometimes specified that the absolute difference between $\bar{X}$ and μ may be at most $100k\%$ of the unknown population standard deviation σ. For a confidence coefficient of 95%, it thus follows that

$$P\left(-\sigma k \leq \bar{X} - \mu \leq \sigma k\right) = 0.95$$

or

$$P\left(-k\sqrt{n} \leq Z \leq k\sqrt{n}\right) = 0.95$$

From Table B.2 $k\sqrt{n} = 1.96$ and therefore

$$n = \left(\frac{1.96}{k}\right)^2$$

For $k = 0.10$ it follows that, for example

$$n = \left(\frac{1.96}{0.10}\right)^2 \approx 384$$

5 Closing remarks

5.1 Two-sample case

It is sometimes important to estimate the difference between the corresponding parameters of two different populations. Thus, for example, it can be important to determine the difference $\mu_1 - \mu_2$ in the mean house prices per square metre of houses in Johannesburg (where a sample of size n_1 was drawn) and Cape Town (where a sample of size n_2 was drawn). For a company marketing dishwashers, it is also important to estimate the difference $\pi_1 - \pi_2$ in the proportions of Johannesburg families and Pretoria families who use dishwashers. If two independent random samples of sizes n_1 and n_2 are drawn from populations with means μ_1 and μ_2 respectively, $\bar{X}_1 - \bar{X}_2$ is a point estimator of the difference $\mu_1 - \mu_2$. Similarly, $P_1 - P_2$ is a point estimator of the difference $\pi_1 - \pi_2$ in proportions. Confidence intervals for $\mu_1 - \mu_2$ and $\pi_1 - \pi_2$ and for $\frac{\sigma_1^2}{\sigma_2^2}$ can be found in a manner analogous to the one-sample case (see Table 11.3).

5.2 Finite correction factor

If the population from which the sample is drawn is relatively small, more accurate approximations for the standard errors of estimation of $\bar{X}$ and P are obtained by multiplying their formulas (see note (b) in Section 2) by a correction factor. For the parameter μ, the correction factor is $\sqrt{\frac{N-n}{N}}$, while for the parameter π, the correction factor is $\sqrt{\frac{N-n}{N-1}}$. The confidence intervals for μ and π then become somewhat shorter (see Table 11.3).

5.3 Table of point estimators and confidence intervals

For the reader's convenience an exposition of all the point estimators and confidence intervals for the situations discussed in this chapter appears in Table 11.3.

Table 11.3 Point estimators and 95% confidence intervals for different unknown population parameters

Parameter	Population size N	Assumptions	Point estimator	95% confidence bounds
μ	Large	(i) Sample from $n(\mu, \sigma^2)$ (ii) n large	$\bar{X}$	(i) $\bar{x} \pm t_{n-1,\,0.025}\frac{s}{\sqrt{n}}$ (ii) $\bar{x} \pm 1.96\frac{s}{\sqrt{n}}$
	Known and relatively small	(iii) Sample from $n(\mu, \sigma^2)$		(iii) $\bar{x} \pm t_{n-1,\,0.025}\left(\frac{s}{\sqrt{n}}\right)\sqrt{\frac{N-n}{N}}$
$\mu_1 - \mu_2$	Large	(i) Independent samples from $n(\mu_1, \sigma^2)$ and $n(\mu_2, \sigma^2)$ populations (ii) Independent samples of sizes n_1 and n_2	$\bar{X}_1 - \bar{X}_2$	(i) $\bar{x}_1 - \bar{x}_2 \pm t_{n_1+n_2-2,\,0.025}\, a\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ with $a = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}}$ (ii) $\bar{x}_1 - \bar{x}_2 \pm 1.96\,b$ with $b = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$
π	Large	(i) n large and π not too close to 0 or 1	$P = \frac{X}{n}$	(i) $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$
	Known and relatively small	(ii) n large and π not too close to 0 or 1		(ii) $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$
$\pi_1 - \pi_2$	Large	n_1 and n_2 both large while π_1 and π_2 not too close to 0 or 1	$P_1 - P_2$ with $P_1 = \frac{X_1}{n_1}$ and $P_2 = \frac{X_2}{n_2}$	$p_1 - p_2 \pm 1.96\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$
σ^2	Large	Sample from $n(\mu, \sigma^2)$ population	S^2	$\left(\frac{(n-1)s^2}{\chi^2_{n-1,\,0.025}}, \frac{(n-1)s^2}{\chi^2_{n-1,\,0.975}}\right)$
$\frac{\sigma_1^2}{\sigma_2^2}$	Large	Independent samples from $n(\mu_1, \sigma_1^2)$ and $n(\mu_2, \sigma_2^2)$ populations	$\frac{S_1^2}{S_2^2}$	$\left(\frac{1}{F_{n_1-1,\,n_2-1;\,0.025}}\frac{s_1^2}{s_2^2}, F_{n_2-1,\,n_1-1;\,0.025}\frac{s_1^2}{s_2^2}\right)$

Exercises

1 Ten new cars of the same model are selected randomly and their fuel consumption is determined by subjecting them to a road test of 500 km. The consumption (ℓ/100 km) is as follows:

8.93	7.75	7.90	8.20	8.41
8.50	8.05	7.93	8.61	8.33

If previous experience has shown that the fuel consumption of cars for the same model is distributed n(μ, σ^2), determine:

(a) Estimates of μ and σ^2;
(b) 95% confidence intervals for μ and σ^2.

2 After the Minister of Finance's budget speech in Parliament, 40 South African economists are asked whether they feel that the budget announced will stimulate economic growth. Fifteen of the economists support the view that this will be the case.

(a) Calculate π, the proportion of economists in South Africa who regard the budget as stimulating for growth.
(b) What is the standard error of estimation if π is estimated by P?
(c) Determine a 95% confidence interval for π.

3 It is known that the maths marks of students are normally distributed with mean μ and standard deviation σ. After 800 first-year students have written a test, sixteen of the test scores are drawn at random. The scores are as follows:

50	82	77	64
58	65	55	53
62	69	70	58
60	36	85	54

(a) Estimate μ and σ^2.
(b) What is the standard error if μ is estimated by $\bar{X}$?
(c) Determine 95% confidence intervals for μ and σ^2.

4 Study the valuations of residences in Durban with numbers 51 to 90 (Appendix A, Data Set 1).

(a) Estimate the mean valuation, μ, of houses in Durban.
(b) Find a 99% confidence interval for μ.
(c) If the difference between $\bar{X}$ and μ with a probability of 0.95 may not exceed R10 000, is this sample of valuations large enough for estimating μ with such accuracy? If not, how many additional observations should be drawn from this populations in order to obtain the required accuracy?

5 Consider the plot sizes of residences in Johannesburg. If all 843 houses referred to in Appendix A, Data Set 1, are regarded as a random sample from the population, determine the following:

(a) An estimation of π, the proportion of residences in Johannesburg with plots smaller than $750m^2$.
(b) The standard error of the estimation in (a).
(c) A 95% and a 99% confidence interval for π.
(d) If it is important for the difference between P and π, with a probability of 0.95, to be no more than 0.03, is the sample size adequate? If not, how many more observations should be drawn from this population before the desired accuracy is obtained?

6 Consider the population given in Example 11.2. By regarding house numbers 701 to 705, 706 to 710, ..., 796 to 800 as twenty random samples of size 5 each, compare the arithmetic average and sample median as estimators of μ, as in Example 11.2.

7 A property developer plans to develop a new suburb in Pretoria. The developer needs to estimate the mean valuation, μ, in the area to within R5 000, with a probability of 95%. The developer has access to Data Set 1 in Appendix A and uses the data of the houses in Pretoria numbered 301 to 330 as the point of departure.
(a) Determine the initial estimate of μ from the above data and
(b) find the sample size required to obtain the specified accuracy.
(c) If this value of n is greater than 30, obtain the additional observations required, namely house numbers 331, 332, ... and calculate the developer's final estimate of μ.
(d) Find a 95% confidence interval for μ.
(e) If the developer needs to estimate the mean land valuation correctly within R2 000 and with a probability of 0.95, calculate the required sample size.
(f) Suppose the developer wants to estimate μ correctly within R5 000, but with a probability of 0.99, what is the required sample size?

8 In the course of three days 210 car accidents occur within a certain municipal area. These accidents are divided into

Fatal (deaths)	Serious injuries	Minor injuries	No injuries	Total
3	12	23	172	210

(a) Estimate the proportion of accidents in the municipal area that (i) resulted in no injuries, and (ii) were fatal.
(b) What are the standard errors of the estimations in (a)? What do you conclude from this?
(c) If the data above is representative of traffic accidents in the area for the whole month, find, for the month, a 95% confidence interval for the proportion of accidents that did not result in injuries.

9 A pharmaceutical company decides to test a new drug that is expected to relieve rheumatism on patients suffering from this condition. This will enable the company to determine what percentage of rheumatism sufferers experience relief with the drug.
   (a) If the percentage has to be determined accurately to within 5% with a probability of 99%, determine the sample size.
   (b) If 150 rheumatism sufferers are involved and 103 find relief, determine a 95% and a 99% confidence interval for the proportion of sufferers who will experience relief with the drug.

10 In a survey of study habits among first-year students at a certain university, 85 first-year students are involved in a random sample. Their average daily study time on weekdays is three and three-quarter hours, with a standard deviation of 45 minutes. Only 25 students usually study for longer than five hours per day.
   (a) Find a 95% confidence interval for the mean number of study hours on weekdays among first-year students at the university.
   (b) Determine a 99% confidence interval for the proportion of students who study more than five hours per day on weekdays.
   (c) If the mean number of study hours on weekdays needs to be determined accurately to within five minutes with a probability of 95%, how many students should be involved in the investigation?
   (d) If study time per day is distributed $n(\mu, \sigma^2)$ and there are 2 320 first-year students at the university, repeat (a) and (b) by using a finite correction factor (see Table 11.3).

11 Fifteen selling prices (R1 000) selected randomly from the Cape Town population of selling prices (Appendix A, Data Set 1) are the following:

150.0	97.0	95.0
177.0	200.0	142.3
147.0	165.0	320.0
165.0	168.9	70.0
130.3	920.0	106.0

Suppose that μ and σ are the mean and standard deviation of selling prices of residences in Cape Town for the period October to December 1991.
   (a) Find point estimates for μ and σ.
   (b) Calculate a 95% confidence interval for μ. (You may assume that the selling prices are normally distributed.)

12 Study the first 100 selling prices of Cape Town residences and assume that π denotes the proportion of residences with a selling price of more than R200 000 in the period October to December 1991.
   (a) Estimate π from the sample of 100 observations.
   (b) Calculate a 95% and a 99% confidence interval for π.

13 The average amount and the standard deviation of the amount charged for a men's haircut at 25 hairdressers in Pretoria is

$\bar{x}$ = R12.70 en s = R1.25.

Determine a 90% and a 95% confidence interval for μ, the mean amount for this service in Pretoria.

14 The board of directors of a large company decides to conduct a survey among its shareholders to determine their views on certain proposed policy changes. There are more than 2 000 shareholders and the directors decide that only 150 randomly selected shareholders will be approached and that their conclusion will be based on the opinions expressed by these shareholders. In the survey, 81 of the 150 shareholders declare themselves in favour of the policy changes and the board decides to implement the changes. Is their action justifiable?

15 During an exam it is noted that of the 404 students writing a paper in a certain room, 38 are left-handed and 92 wear glasses.
(a) Determine point estimators of the percentage of students who are left-handed and also the percentage of students who wear glasses during an exam.
(b) Find 95% confidence intervals for the proportions defined in (a).

16 A sample of size n is drawn from a normal population. If it is required that $|\bar{x} - \mu| \leq e\mu$ with a probability of 0.95, show that

$$n = \left(\frac{1.96\,\tau}{100\,e}\right)^2$$

where $\tau = \frac{\sigma}{\mu} \times 100$ is the population coefficient of variation of X. (Clue: study the deduction in Section 4.1).

17 Suppose it is known that the population coefficient of variation of the heights of adult males is 5%. An anthropologist wants to estimate the mean height μ of adult males of a primitive South American tribe. How large does the sample have to be if the absolute difference between $\bar{X}$ and μ may at most be
(a) 20% of the unknown population standard deviation σ, and
(b) 2% of the population mean μ?

18 The manager of a cinema wants to schedule shows so that a maximum number of shows can be fitted in per day. All shows have to be scheduled between 10:00 and 24:00, and all shows have a supporting program and an interval with a joint duration of 35 minutes. Schedule the starting times for shows if the most recent film guide, which discusses 107 films, gives the average length of films as $\bar{x}$ = 103 minutes. The standard deviation of film length is s = 5 minutes.

19 A tourist guidebook for a certain city cites the mean annual rainfall for the city as 750 mm. However, rainfall figures for 1940 to 1985 show that in 36 of these years it rained less than 750 mm. Would you say that the tourist guidebook is an unreliable source of information with respect to the rainfall?

CHAPTER 12

Hypothesis testing

1 Introduction

Decisions often have to be made in situations where accurate information is not available. In some cases it is not possible to obtain full information because the cost involved is too high or because the process of collecting the information takes too long. In other cases, for instance where the life expectancy of parts is studied, it is not practically possible to obtain full information because the parts are destroyed in the measuring process.

In practice, the decision maker who has little information, often has to make a choice between two decisions, say B_0 and B_1. If the choice is made on the basis of information obtained from a sample, it is very important that the decision be made on a basis that is scientifically accountable.

EXAMPLE 12.1

All newly appointed representatives at a certain firm are required to complete a two-week compulsory training course in sales techniques. The course has been presented for ten years at a current cost of R3 000 per person. A consultant approaches the firm and offers to train newly appointed staff with the aid of a modern course in sales techniques at a total estimated cost of R5 000 per person. This course also lasts two weeks. The board of directors decides to accept the offer, provided the mean first-year sales of representatives who complete the more expensive course are **significantly** higher than the mean first-year sales of those who completed the traditional course.

Because the old training course has been running for ten years, the firm has, among other things, the sales figures for each representative for the first year after completion of the course. With due consideration to the depreciation of the rand, these mean first-year sales are translated into a current money value of R300 000. Obviously, the firm does not have a corresponding figure available for sales after the new course. For this reason ten persons are selected at random from all the new appointments to be trained by the consultant. After training, the mean first-year sales of the selected ten can then be compared with the figure of R300 000.

In the above situation, it is clear that the directors have to make a choice between the following options:

B_0: The mean first-year sales of persons who have undergone the new training is also R300 000. The status quo is therefore maintained.

B_1: The mean first-year sales of those who have completed the new course are significantly higher than R300 000. The consultant's offer is therefore accepted.

Obviously the choice between the two options will be based on the information obtained from the random sample of size 10.

In estimation theory as discussed in Chapter 11, we are concerned with the estimates of unknown population parameters, for instance π, the proportion of individuals in the population with a certain characteristic, and μ, the population mean. These estimates are based on observed random samples. In contrast, hypothesis testing usually makes certain **suppositions** about the value of an unknown population parameter. If a choice has to be made between one of two decisions, each is usually expressed in the form of a **hypothesis** (supposition). It is important to note that one of these hypotheses should always be formulated in such a way that the value of the population parameter concerned is uniquely specified.

A hypothesis that uniquely specifies the population parameter concerned, is called a **null hypothesis** and is indicated by the symbol H_0.

The **alternative hypothesis**, indicated by H_1, specifies for the population parameter a set of values that is not specified by the null hypothesis and that is important to the specific problem.

EXAMPLE 12.2

Suppose the first-year sales figures for the ten representatives who completed the new training course are taken from a population with unknown mean first-year sales, μ. The firm's decision is then based on a choice between the following hypotheses:

H_0: $\mu = 300\,000$

H_1: $\mu > 300\,000$

In the section on estimation theory we indicated that $\bar{x}$ can be used as an estimate of μ. Basically, the choice between H_0 and H_1 depends on the size of the difference between the observed arithmetic mean $\bar{x}$ and 300 000. As $\bar{X}$ is a continuous stochastic variable, the observed $\bar{x}$ would seldom be equal to the unknown population mean μ. For this reason H_0 is accepted if the difference between $\bar{x}$ and the μ value specified by H_0 does not exceed a **predetermined value**. Otherwise H_1 is accepted. The determination of this value is discussed in Section 3 of this chapter.

Notes

(a) We distinguish between two types of hypotheses, namely those that specify only the value(s) of unknown population parameters and those that specify the nature of the distribution of the relevant stochastic variable as, for example, the hypothesis that the stochastic variable X has a normal distribution. In this chapter we focus only on the former, while the latter is discussed in Chapter 15.

(b) The null hypothesis in Example 12.2 specifies a single value of μ and is therefore called a **simple hypothesis**, while the alternative hypothesis specifies a set of possible values and is known as a **composite hypothesis.**

2 Statistical hypotheses

The process of decision making, as discussed in this chapter, is known as statistical hypothesis testing.

> The technique used to make a decision on the correctness of a **hypothesis** on the basis of information obtained from an observed sample, is known as **hypothesis testing**.

As mentioned in Section 1, we distinguish between the **null hypothesis** and the **alternative hypothesis** and our current discussion is limited to hypothesis testing with regard to a single parameter of the distribution from which the sample was selected.

EXAMPLE 12.3

Suppose the current unemployment figure is said to be higher than 30%. Because existing census data could be obsolete, a representative sample of 2 500 persons from the section of the population that is regarded as potentially economically active, is investigated. Define the stochastic variable X as the number of respondents who indicate that they are unemployed. From the theory discussed in Chapter 10 it follows that X is distributed as bi(2 500, π) and the hypothesis to be tested is therefore the hypothesis

$$H_0: \pi = 0.30$$

The appropriate alternative in this case is

$$H_1: \pi > 0.30$$

When testing a statistical hypothesis, H_0 and H_1 are formulated in such a way that the acceptance of H_0 or H_1 implies decision on one of two contrasting options. As far as possible, the null hypothesis is always simple and the alternative is usually a composite hypothesis. We encounter, for instance, the following situations:

(a) A psychologist maintains that the population mean IQ of twelve-year-old children is not 100. Such a statement can be tested by formulating the hypotheses as follows:

H_0: $\mu = 100$

H_1: $\mu \neq 100$

(b) A health inspector alleges that π, the proportion of the population of dairy herds in the Transvaal infected by brucellosis, is higher than 10%. The inspector's allegation can be formulated as an upper-tailed hypothesis testing problem:

H_0: $\pi = 0.10$

H_1: $\pi > 0.10$

(c) At a steel plant chrome metal is plated onto steel plating. The customer requires the standard deviation in chrome thickness to be less than 15 mg/m^2. The requirement is formulated as follows in a lower-tailed hypothesis testing problem:

H_0: $\sigma^2 = (15)^2$

H_1: $\sigma^2 < (15)^2$

As can be seen from these three examples, we distinguish between **two-tailed** (or **two-sided**) testing (case (a)), **upper-tailed** (or **upper one-sided**) testing (case (b)) and **lower-tailed** (or **lower one-sided**) testing (case (c)).

EXAMPLE 12.4

Suppose face bricks are manufactured at a certain brickyard, and some do not comply with the quality specifications. A construction firm that uses these bricks is prepared to accept a consignment from the brickyard, provided π, the population proportion of rejected bricks, is, at most, 0.05. The firm's planned strategy can be formulated as follows as an upper-tailed hypothesis testing problem:

H_0: $\pi = 0.05$

H_1: $\pi > 0.05$

Before a decision can be taken, the quality of a random sample of face bricks must be examined. If H_1 is accepted, the firm concludes that the proportion of face bricks of poor quality is too high, and the consignment is therefore not accepted. On the other hand, the acceptance of H_0 does not imply that π is exactly 0.05. The actual value of π could be considerably less than 0.05, but as this can only be to the advantage of the construction firm, acceptance of such a consignment would not be a wrong decision.

Of course, π can also be considerably greater than 0.05, even though this is not reflected in the value of p, the observed sample ratio of bricks of poor quality.

Despite this, we do not, in such a case, reject H_0, since the difference between p and the supposed value of π is not sufficient to reject H_0. Essentially, the acceptance of H_0 means that the difference between the supposed value of an unknown population parameter and the corresponding estimate of this parameter on the basis of a sample is not large enough to cause H_0 to be rejected. It is therefore important to note that we should avoid the terminology 'H_0 is accepted' and rather say that H_0 cannot be rejected.

3 Errors and risks

To test hypotheses regarding the unknown population parameters μ, σ^2 and π, we use the fact that $\bar{X}$, S^2 and P are the respective estimators of these parameters. From the unknown population parameters and their estimators, the so-called pivotal quantities are formed (see Chapter 10) and the distributions of the pivotal quantities are used in the decision-making process. Thus, for example, it follows from Chapter 10, Section 10 that

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \text{ has a } t(n-1) \text{ distribution,}$$

$$Y = \frac{(n-1)S^2}{\sigma^2} \text{ a } \chi^2(n-1) \text{ distribution, and}$$

$$Z = \frac{P - \pi}{\sqrt{\pi(1-\pi)/n}} \text{ approximately a } n(0,1) \text{ distribution}$$

provided that, in each case, the assumptions are complied with.

As the pivotal quantities above reveal, the formulation of the null hypothesis means that the distribution of the pivotal quantity concerned is specified in full. Thus, for example, for a random sample from a $n(\mu, \sigma^2)$ population with $H_0: \mu = 100$ and $H_1: \mu \neq 100$,

$$T = \frac{\bar{X} - 100}{S/\sqrt{n}}$$

has a $t(n-1)$ distribution, provided H_0 is true. If the observed values of $\bar{X}$ and S, namely $\bar{x}$ and s, are also replaced in the pivotal quantity, the corresponding value of T, namely

$$t = \frac{\bar{x} - 100}{s/\sqrt{n}}$$

will be an observation of the $t(n-1)$ distribution. As soon as the difference between $\bar{x}$ and 100 exceeds a previously selected value, H_0 is rejected. It is clear that the decision to reject H_0 amounts to a comparison of the observed t value with a critical value, a.

Suppose we use the following decision rule or statistical test:

H_0 is rejected if $t \geq a$ or $t \leq -a$, that is if $|t| \geq a$;

H_0 is not rejected if $-a < t < a$, that is if $|t| < a$

The above values $-a$ and a are the critical values and the area to the left of $-a$ and to the right of a is known as the **rejection area** for H_0. Figure 12.1 illustrates this concept.

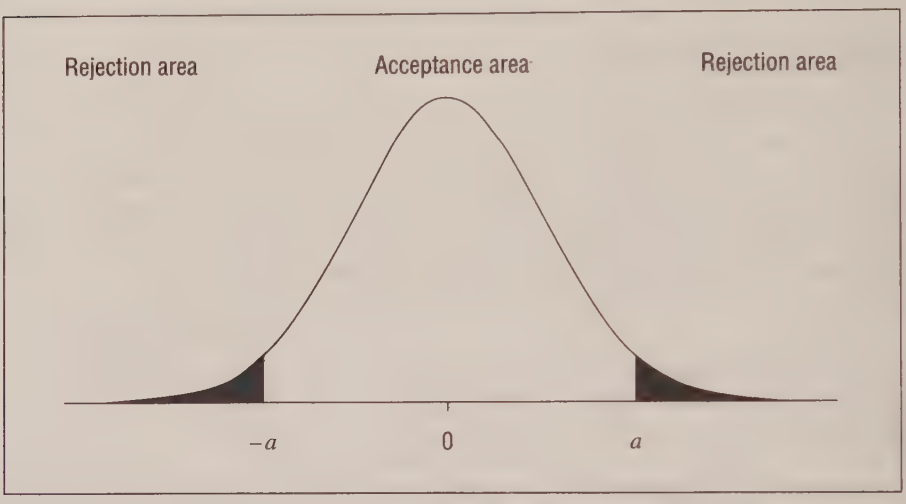


Figure 12.1 Critical values and rejection area for a two-tailed alternative

Since a pivotal quantity is a stochastic variable, we can hardly expect to make the correct decision frequently if it is based on an observed value of the stochastic variable. In the above decision rule, for example, H_0 is rejected if t lies in the rejection area. Theoretically it is possible for t to lie in this area, despite the fact that $\mu = 100$. However, the critical value a is selected in such a way that the probability of this happening is small. A number of outliers in the observed sample from a population with $\mu = 100$ could, for instance, lead to a large $\bar{x}$ value, which would then cause t to fall in the rejection area for H_0. However, the opposite of this situation can also occur; the observed t value can lie in $(-a, a)$, despite the fact that μ is, for instance, 120.

Hypothesis testing on the basis of the observed sample can therefore lead to two divergent types of errors:

I: H_1 is accepted while H_0 is true, and

II: H_0 is accepted while H_1 is true.

These two errors occur due to stochastic variation in the sampling process and cannot be attributed to physical calculation errors.

The erroneous decision to reject H_0 (and therefore to accept H_1 if H_0 is true) is known as the **type I error.**

The erroneous decision to accept H_0 if H_1 is true, is known as the **type II error**.

Table 12.1 contains a summary of the possible decisions we could make in hypothesis testing, as well as the implications of each decision.

Table 12.1 Possible decisions and their implications with regard to hypothesis testing

Decision based on sample	True state of affairs	
	H_0 true	H_1 true
B_0: Accept H_0	Correct decision	Type II error
B_1: Reject H_0	Type I error	Correct decision

The probability of a type I or a type II error occurring is known as the risk associated with hypothesis testing.

The maximum permissable risk associated with a type I error is known as the **level of significance**, α.

For the sake of brevity we indicate a type I error by

$$(H_0 \text{ rejected} \mid H_0 \text{ is true})$$

and a type II error by

$$(H_0 \text{ accepted} \mid H_1 \text{ is true})$$

The level of significance, α, can therefore be expressed as follows:

$$\alpha = P(\text{type I error})$$
$$= P(H_0 \text{ rejected} \mid H_0)$$

while we usually use β to indicate the risk associated with a type II error:

$$\beta = P(\text{type II error})$$
$$= P(H_0 \text{ accepted} \mid H_1)$$

From this it follows that the probabilities of correct decisions are given by

$$1 - \alpha = P(H_0 \text{ accepted} \mid H_0)$$
$$1 - \beta = P(H_0 \text{ rejected} \mid H_1)$$

A decision maker would obviously like to limit the α and β risks as much as possible. It can be shown that, for a fixed sample size n, the value of β will increase if α decreases, and vice versa, so that the two risks for fixed n cannot be decreased arbitrarily. However, as it is possible to keep both α and β low if the sample size n increases, the ideal is to specify the α and β values when a survey is being planned and then to find the sample size n in which both specifications are met. This task can be facilitated by the use of operating characteristic curves.

The **operating characteristic curve** for testing a null hypothesis regarding the population mean μ is a graphic representation of $\beta(\mu) = P(H_0 \text{ accepted} \mid \mu)$ at different values for μ at a fixed value for n.

Figure 12.2 shows a typical operating characteristic curve for testing

$H_0: \mu = 100$ against $H_1: \mu < 100$

with a fixed sample size of $n = 30$. The ideal operating characteristic curve would naturally be the curve for which

$$\beta = \begin{cases} 0, \text{ if } \mu < 100 \\ 1, \text{ if } \mu \geq 100. \end{cases}$$

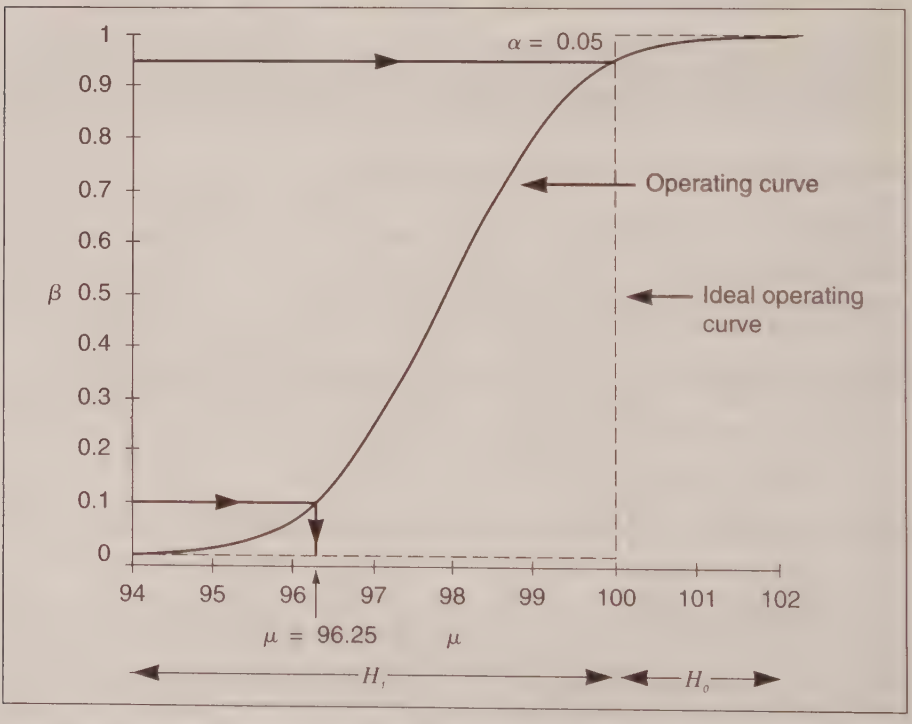


Figure 12.2 Operating characteristic curve and ideal operating curve for testing $H_0: \mu = 100$ against $H_1: \mu < 100$, with $n = 30$

The ideal operating characteristic curve is indicated by a dotted line in the figure. Suppose the decision maker is prepared to take the following risks:

$$\alpha = P(\text{type I error}) = 0.05$$

and

$$\beta = P(\text{type II error}) = 0.10$$

From Figure 12.2 it seems that

$$\text{maximum } \{P(H_0 \text{ accepted} \mid \mu < 96.25)\} = 0.10.$$

If the decision maker specifies both types of error, there must be an indication of the maximum permissible deviation for which he is still prepared to accept H_0. Suppose, for example, that the characteristic being examined is the breaking-strength of steel wire. If 99 is the minimum acceptable breaking-strength, it follows for a sample of size 30 from Figure 12.2 that $\beta \approx 0.80$, so the risk involved in a type II error is far greater than the required 0.10. The only way of complying with all the requirements, is thus to increase the sample size. The probability of H_0 being rejected given that H_1 is true, namely $1 - \beta$, is known as the **power** of the test. In Section 8 of this chapter the operating curves and the calculation of the power functions are discussed further.

If only the α risk is specified, α is usually taken to be 0.05 or 0.01. It has become customary to refer to the cases for which H_0 is rejected as **significant** at a 5% level of significance, and to the cases where H_0 is rejected at a 1% level of significance as **highly significant**.

In situations where β is not specified beforehand, the decision maker should formulate the null and alternative hypotheses in such a way that the incorrect acceptance of H_0 does not entail serious costs or other similar complications. The following example illustrates such a situation.

EXAMPLE 12.5

Suppose a firm with a fleet of vehicles of the same make has been using 'Hold and Cling' tyres for their cars for some time. Given the circumstances in which these vehicles are used, the tyres achieve an average distance of 40 000 km. The firm is highly satisfied with this type of tyre, but receives an offer from the manufacturers of 'Light Tread' tyres for tyres at a higher discount.

The manufacturers of this tyre claim that their product achieves a distance of more than 40 000 km. The firm concerned decides to test this claim at a 1% level of significance. A sample of size 20 is decided on. The firm now has to choose between the following two options:

B_0: Continue with 'Hold and Cling'.

B_1: Change to 'Light Tread'.

Consider the following two possible approaches:

(i) In terms of hypotheses we can formulate the decisions as follows:

H_0: $\mu = 40\,000$

H_1: $\mu > 40\,000$

where μ is the mean distance of 'Light Tread' tyres. In this case 'Light Tread' tyres will be used once the sample mean distance of 'Light Tread' tyres is significantly higher than 40 000 km. If H_0 is not rejected, the firm chooses option B_0. If this decision is incorrect, a type II error has been made. Although it would be beneficial to change to 'Light Tread' tyres, the implication of the incorrect acceptance of B_0 is not serious, because the firm has always been satisfied with 'Hold and Cling'.

(ii) On the other hand, suppose the hypotheses were formulated as follows:

H_0: $\mu = 40\,000$

H_1: $\mu < 40\,000$

If H_0 is not rejected in this case, the company will change to 'Light Tread' tyres, that is, option B_1 is chosen. Suppose the sample results lead to H_0 not being rejected, while μ is actually 30 000 km. The type II error now committed could possibly result in costs for which the higher discount would hardly compensate.

It is clear from the above discussion that approach (i) makes far stricter demands on 'Light Tread' tyres than approach (ii), and that the former approach is safer for the fleet owner.

4 Test statistics and decision rules

Consider again testing of

H_0: $\mu = 100$ against H_1: $\mu \neq 100$

on the basis of a random sample of size n from a $n(\mu, \sigma^2)$ population. As at the beginning of Section 3, we note that the choice between H_0 and H_1 is dependent upon the observed value of the pivotal quantity at which the H_0 value of the parameter for the pivotal quantity was set. The decision function that is established if the H_0 value of the relevant parameter is substituted in the pivotal quantity, is known as the test statistic.

A **test statistic** is a function of the null hypothesis and the random sample $X_1, X_2, \ldots, X_n$ and is used to decide between H_0 and H_1. Under H_0 the sampling distribution of the test statistic is fully specified.

Note

The symbols μ_0, σ_0^2 and π_0 will be used from now on to indicate the H_0-values of μ, σ^2 and π respectively. For example,

$$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

is the test statistic for H_0: $\mu = \mu_0$

EXAMPLE 12.6

Suppose an estate agent maintains that the population mean selling price, μ, of residences in Johannesburg from October to December 1991 was R180 000. Consider the test of the hypothesis

H_0: $\mu = 180\,000$ rand against

H_1: $\mu \neq 180\,000$ rand.

As test statistic we use the function

$$Z = \frac{\bar{X} - 180\,000}{S/\sqrt{n}}$$

which approximately has a n(0, 1) distribution under H_0 for large values. This result follows from the use of the central limit theorem. According to this theorem the mean of a random sample $X_1, X_2, \ldots, X_n$ from a population with a mean μ and standard deviation σ is approximately normally distributed, provided n is large.

The fact that the sampling distribution of the test statistic is fully specified under H_0, provides us with a criterion according to which a choice between H_0 and H_1 can be made. Suppose, for example, that the observed $\bar{x}$ is considerably lower than R180 000 and that z is therefore considerably lower than 0. To accept H_0 in this case means that we have to assume that this $\bar{x}$ is an extreme value from a distribution with mean μ = R180 000. However, such large differences cannot reasonably be attributed time and again to stochastic variation. It is more probable that $\bar{x}$ is an observation from a distribution with μ considerably lower than R180 000. In the case described above the sample result will lead to questioning the correctness of H_0 and the consequent acceptance of H_1. On the other hand, if $\bar{x}$ differs from R180 000 by only a small amount and z therefore shows only a minor difference from 0, it is quite possible that this difference can be attributed to stochastic variation. In this case our decision will be not to reject H_0.

Since Z approximately has a n(0, 1) distribution under H_0 for $n = 100$, 95% of all observed z values can be expected to lie between the limits –1.96 and 1.96. An observed value of, for example, $z = -1.03$ can therefore reasonably be attributed to stochastic variation. However, suppose that $\bar{x}$ differs from R180 000 to the extent that $z = -2.41$, then we can

(a) regard this value as an extreme value from a n(0, 1) distribution or

(b) assume that the real population mean differs from the supposed value of R180 000.

As only 5% of all n(0, 1) observations lie outside the interval (–1.96, 1.96), (a) is unlikely and we consequently decide on (b), that is, we decide that $\mu \neq$ R180 000.

The procedure followed above can be summarized as follows:

H_0 is rejected if $z \leq -1.96$ or $z \geq 1.96$.

H_0 is not rejected if $-1.96 < z < 1.96$.

Note that in many similar situations, in 5% of the cases, according to this procedure we would reject H_0 where H_0 is actually true, since 5% of the observed z values can be expected to lie outside the interval (–1.96, 1.96) if $\mu =$ R180 000. This value of 5% is naturally the level of significance or the α risk already defined, while –1.96 and 1.96 are the critical values discussed in the previous section. In Example 12.7 we solve this specific problem.

From Example 12.6 it is clear that with hypothesis testing there are two mutually exclusive sets of observed test statistic values:

(a) The set of values for which H_0 is not rejected.

(b) The set of values for which H_0 is rejected.

The value (or values) separating the area in which H_0 is not rejected from the area in which H_0 is rejected, is known as the **critical value** (or values) of the test.

If the level of significance α is specified, the critical value(s) can be read from the appropriate statistical table. Figure 12.3 shows the critical values for test statistics that have the n(0, 1), the t$(n-1)$, the $\chi^2(n-1)$ and the F(n_1-1, n_2-1) distributions, where it is assumed that $\alpha = 0.05$.

We are now in a position to provide an explanation of the different steps to be followed in testing a hypothesis.

4.1 Steps in hypothesis testing

Step 1

Formulate the null hypothesis and alternative hypothesis to correspond with the problem concerned.

Step 2

Specify the level of significance α. Decide on the appropriate test statistic and indicate its distribution under H_0.

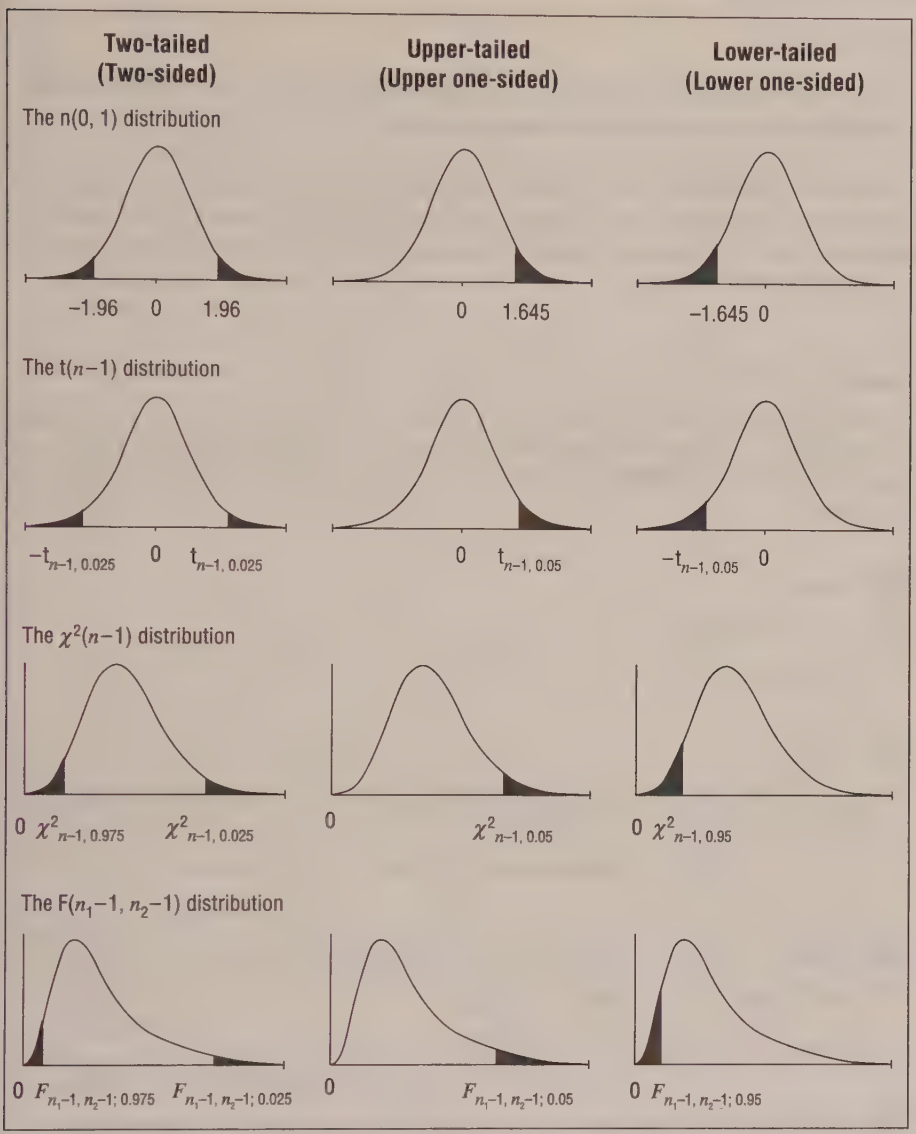


Figure 12.3 Critical values of test statistics that are distributed respectively n(0, 1), t($n-1$), $\chi^2(n-1)$ and F(n_1-1, n_2-1) ($\alpha = 0.05$)

Step 3

Give the decision rule. This step involves the choice of a critical value or values from Appendix B. Figure 12.3 is also useful here.

Step 4

Calculate the observed value of the test statistic.

Step 5

Reject H_0 or decide not to reject H_0 in accordance with the decision rule. Draw a conclusion in terms of the problem concerned.

It must be emphasized that H_0 and H_1 should not be formulated on the basis of conclusions drawn from the observed data. The choice of hypotheses is suggested by the problem being investigated and should therefore be made before collecting the data.

EXAMPLE 12.7

To illustrate the previous five steps, we consider Example 12.6, where μ, the mean selling price of residences in Johannesburg during October to December 1991, is of relevance. The null hypothesis will be tested on the basis of a random sample of size 100.

Step 1

$H_0: \mu = 180\,000$

$H_1: \mu \neq 180\,000$

Step 2

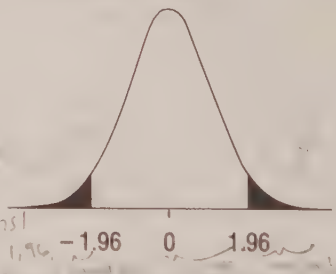
Let $\alpha = 0.05$. Under H_0 and by using the central limit theorem the test statistic

$$Z = \frac{\bar{X} - 180\,000}{S/\sqrt{100}}$$ approximately has a n(0, 1) distribution.

Step 3

H_0 is rejected if $z \leq -1.96$ or $z \geq 1.96$

H_0 is not rejected if $-1.96 < z < 1.96$



Step 4

For a random sample of 100 selling prices it was found that $\bar{x} = \text{R}158\,568$ and $s = \text{R}82\,106$. The observed value of the test statistic is therefore

$$z = \frac{158\,568 - 180\,000}{82\,106/\sqrt{100}}$$

$$= -2.61$$

Step 5

Since $z = -2.61 < -1.96$, H_0 is rejected. In terms of the problem, the conclusion is that the mean selling price of residences in Johannesburg during the period implicated is not R180 000.

Notes

(a) It is important to note that in the above example we did not conclude that μ is smaller than R180 000, but only that μ is not equal to R180 000. The reason for this is that the alternative hypothesis selected is a two-tailed alternative. It is wrong to adjust the alternative hypothesis to the result of a statistical test. If necessary, H_0 and H_1 can be formulated differently in a follow-up investigation.

(b) As an alternative to the above method of testing hypotheses, a confidence interval for μ can be constructed (see Chapter 11). If μ_0 is an element of this interval, H_0 is not rejected. The level of significance α determines the confidence coefficient of the confidence interval. For example, the confidence coefficient in the above situation should be taken as 95%, as $\alpha = 0.05$. The interval is given by

$$\left(\bar{x} - \frac{1.96s}{\sqrt{100}}, \bar{x} + \frac{1.96s}{\sqrt{100}}\right)$$

$$= \left(158\,568 - \frac{1.96(82\,106)}{\sqrt{100}}, 158\,568 + \frac{1.96(82\,106)}{\sqrt{100}}\right)$$

$$= (142\,475.22, 174\,660.78)$$

$$\approx (142\,475, 174\,661)$$

As the μ_0 value, namely 180 000, lies outside this interval, H_0 is rejected as before.

(c) The five steps followed above (and which can be adjusted slightly if a confidence interval is calculated) form the basis of hypothesis testing with respect to the parameters of distributions. In section 5 hypotheses in the form of

$$H_0: \mu = \mu_0;\ H_0: \pi = \pi_0 \text{ and } H_0: \sigma^2 = \sigma_0^2$$

are considered. This is known as the **one-sample case**. In Section 6 hypotheses in the form of

$$H_0: \mu_1 = \mu_2;\ H_0: \pi_1 = \pi_2 \text{ and } H_0: \sigma_1^2 = \sigma_2^2$$

are tested. These all form part of the so-called **two-sample case**.

4.2 *p*-values in hypothesis testing

In hypothesis testing problems, the observed value of the test statistic is compared with a critical value. Take, for example, testing $H_0: \mu = \mu_0$ against $H_1: \mu > \mu_0$ with the aid of the test statistic Z, which has a n(0, 1) distribution where n is large. If $\alpha = 0.01$, the critical value $z_{0.01} = 2.326$. Now suppose

that the observed test statistic value z equals 2.87, then H_0 is rejected. For $z = 3.43$, for example, H_0 is also rejected.

At present, a large variety of computer programs can do statistical inference. The output of most of these programs includes a p-value that provides the user with additional information which will determine the degree of confidence with which the null hypothesis can be rejected or accepted. When using p-values, the following rule applies:

> Reject H_0 if the p-value $\leq \alpha$.
> Do not reject H_0 if the p-value $> \alpha$.

The p-value for the hypothesis testing problem mentioned above is calculated as follows:

$z = 2.87$ so that the p-value $= P(Z \geq 2.87) = 0.5 - 0.49795 = 0.00205$

$z = 3.43$ so that the p-value $= P(Z \geq 3.43) = 0.5 - 0.49970 = 0.00003$

Since the p-value for $z = 3.43$ is lower than the p-value for $z = 2.87$, we can reject H_0 with greater confidence if $z = 3.43$ compared to $z = 2.87$.

> The ***p*-value** is the probability that the test statistic under H_0 is equal to the observed value of the test statistic, or is more extreme in the direction suggested by the alternative. It is also known as the **exceedance probability**.

The appropriate formulas for calculating p-values will be given in the various examples (see for instance Table 12.2). Thus it follows from Table B.2 for Example 12.7 that the p-value $= 2P(T \geq |-2.61|) = 2[0.5 - 0.49547] = 0.00906$. Since $0.00906 < \alpha$, H_0 is rejected.

5 The one-sample case

In this section we look at different cases of hypothesis testing with regard to a single population parameter. In each case, the various assumptions that are relevant when hypotheses with regard to μ, π and σ^2 respectively are tested, will be mentioned.

5.1 The population mean μ

The population mean μ is definitely the most important parameter of a distribution and we often require information about the value of μ or we want to determine whether or not μ has changed due to external factors. If, for example, a new production technique is developed it is important to know whether or not the mean manufacturing time per part has changed. Suppose a

new drug for treating a certain strain of influenza is marketed. The question then arises as to whether or not the mean recovery time for patients with this illness is decreased when the new drug replaces the best existing drug. An insurance company can, for example, try to determine whether the life expectancy of fifty-year old women in the RSA has remained unchanged since the 1991 census. Each of these questions can be formulated in terms of a hypothesis testing problem with null hypothesis H_0: $\mu = \mu_0$ and a suitable alternative hypothesis.

When testing hypotheses about μ we usually distinguish between the following two cases:

(a) $X_1, X_2, \ldots, X_n$ is a random sample from a population with mean μ and variance σ^2. The sample is large enough to apply the central limit theorem and it follows that

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}} \text{ approximately has a n}(0, 1) \text{ distribution.}$$

(b) $X_1, X_2, \ldots, X_n$ is a sample from a n(μ, σ^2) population and therefore

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \text{ has the t}(n - 1) \text{ distribution.}$$

This result applies if $n \geq 2$. In order to distinguish between the two cases, we usually refer to (a) as the **large-sample case** and to (b) as the **small-sample case**.

5.1.1 Large-sample test for H_0: $\mu = \mu_0$

Table 12.2 contains a summary of the test procedure, where it is accepted that the random sample is large enough to apply the central limit theorem.

Table 12.2 Testing H_0: $\mu = \mu_0$ for large n

Test statistic	Distribution of test statistic under H_0	Table with critical values
$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	Approximately n(0, 1)	Table B.2
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
H_1: $\mu \neq \mu_0$	$z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$	$2P(Z \geq \lvert z \rvert)$
H_1: $\mu > \mu_0$	$z \geq z_\alpha$	$P(Z \geq z)$
H_1: $\mu < \mu_0$	$z \leq -z_\alpha$	$P(Z \leq z)$

$z_{0.05} = 1.645$ $z_{0.025} = 1.96$ $z_{0.01} = 2.326$ $z_{0.005} = 2.576$

EXAMPLE 12.8

A clothing retailer is of the opinion that customers buying on account can be encouraged to buy more if they are offered a 10% discount. This offer is made to a random sample of 200 customers who have spent R1 000 or more during the past year. After a year the sample shows a mean increase in purchases of R372.14, with a standard deviation of R489.03. The retailer intends expanding the 10% discount to all account holders if they can be sure (1% level of significance) that the mean increase in purchases is significantly higher than R250. The problem is solved as follows by means of hypothesis testing:

Step 1

H_0: $\mu = 250$

H_0: $\mu > 250$

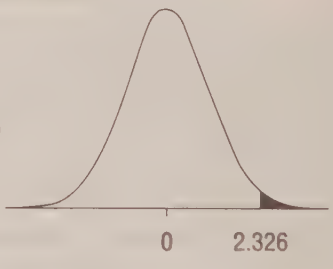
Step 2

Let $\alpha = 0.01$. In this case $n = 200$, so that under H_0 the test statistic

$$Z = \frac{\bar{X} - 250}{S/\sqrt{200}}$$ approximately has a n(0, 1) distribution.

Step 3

Reject H_0 if $z \geq 2.326$ (see accompanying sketch)
or if the p-value $= P(Z \geq z) \leq 0.01$



Step 4

From the data it follows that $\bar{x} =$ R372.14 and $s =$ R489.03.

Therefore $z = \frac{372.14 - 250}{489.03/\sqrt{200}} = 3.53$

Step 5

Since $z > 2.326$, H_0 is rejected. Alternatively the p-value can be calculated as follows using Table B.2: p-value $= P(Z \geq z) = P(Z \geq 3.53) = (0.5 - 0.49979) = 0.00021$. Since this p-value is much lower than $\alpha = 0.01$, H_0 is rejected. We thus have a highly significant result and can therefore confidently conclude that μ, the population mean increase in sales, is significantly higher than R250.

5.1.2 Small-sample test for H_0: $\mu = \mu_0$

If hypotheses on the population mean are tested and the sample is small, and if we can assume that the random sample $X_1, X_2, \ldots, X_n$ is taken from a n(μ, σ^2) population, the test procedure summarized in Table 12.3 may be used.

Table 12.3 Testing H_0: $\mu = \mu_0$ for n small and with observations from a normal distribution

Test statistic	Distribution of test statistic under H_0	Table with critical values
$T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$	t($n-1$)	Table B.3
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
$H_1: \mu \neq \mu_0$	$t \leq -t_{n-1,\,\alpha/2}$ or $t \geq t_{n-1,\,\alpha/2}$	$2P(T \geq \lvert t \rvert)$
$H_1: \mu > \mu_0$	$t \geq t_{n-1,\,\alpha}$	$P(T \geq t)$
$H_1: \mu < \mu_0$	$t \leq -t_{n-1,\,\alpha}$	$P(T \leq t)$

EXAMPLE 12.9

In Example 12.1 a firm's board of directors has to decide whether newly appointed representatives would take the traditional course in sales techniques or, would instead, change to a new course offered by a consultant. Suppose the first-year sales figures of ten representatives selected at random, who completed the new course are as follows:

R287 900	R419 400	R338 300	R287 500	R310 850
R292 600	R390 050	R369 850	R430 400	R338 450

It follows that $\bar{x}$ = R346 530 and s = R53 767.58. As seen in Example 12.2, the firm's decision is based on a choice between the following two hypotheses:

H_0: μ = 300 000 against H_1: μ > 300 000

with 300 000 the mean first-year sales of representatives who completed the traditional course.

Step 1

H_0: μ = 300 000

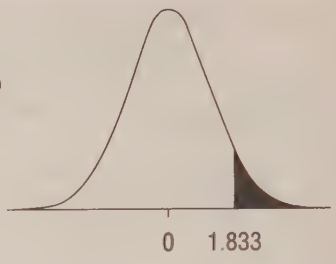
H_1: μ > 300 000

Step 2

Let α = 0.05. Under H_0, the test statistic $T = \dfrac{\bar{X} - 300\,000}{S/\sqrt{10}}$ has the t(9) distribution, provided we can assume that the ten observations are taken from a normal population.

Step 3

Reject H_0 if $t \geq 1.833$ (see accompanying sketch) or if the p-value $= P(T \geq t) \leq 0.05$.



Step 4

$$t = \frac{346\,530 - 300\,000}{53\,767.58/\sqrt{10}}$$

$$= 2.74$$

Step 5

Since $t > 1.833$, H_0 is rejected and we conclude that the mean first-year sales of the ten representatives who attended the new course is significantly higher than R300 000. The firm will consequently change to the new course. Using a computer program, it was calculated that the p-value $= P(T \geq 2.74) = 0.01142$. This p-value is lower than 0.05 and H_0 is therefore rejected. However, had α been chosen as 0.01, it follows that we would not have rejected the null hypothesis.

5.2 The population proportion π

It is often possible to classify the outcome of a stochastic experiment as belonging to one of two mutually exclusive categories. Thus, for example, we can classify persons on the basis of gender as either male or female, a manufactured article on the basis of quality as either defective or non-defective and a house on the basis of roof construction as either flat-roofed or pitched-roofed.

Table 12.4 Testing H_0: $\pi = \pi_0$, on the assumption that the sampling proportion P is approximately normally distributed

Test statistic	Distribution of test statistic under H_0	Table with critical values
$Z = \dfrac{P - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$	n large: approximately n(0, 1)	Tabel B.2
Alternative hypothesis	**Decision rule: Reject H_0 if**	**p-value**
H_1: $\pi \neq \pi_0$	$z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$	$2P(Z \geq \lvert z \rvert)$
H_1: $\pi > \pi_0$	$z \geq z_\alpha$	$P(Z \geq z)$
H_1: $\pi < \pi_0$	$z \leq -z_\alpha$	$P(Z \leq z)$
$z_{0.05} = 1.645$ $z_{0.025} = 1.96$ $z_{0.01} = 2.326$ $z_{0.005} = 2.576$		

As indicated in Chapter 10, such situations involve binomial experiments. With data of this nature it is usually important to confirm an assumption concerning the population proportion of successes, that is, to test the hypothesis

$$H_0: \pi = \pi_0$$

against a suitable alternative. Table 12.4 contains a summary of the test procedure, in which we assume that the required conditions for the normal approach to the binomial distribution have been met (see Chapter 10, Section 8).

EXAMPLE 12.10

An advertising company claims that the television commercial they have created to inform the public about a three-point plan for preventing AIDS is so effective that 75% of the viewers who have seen the commercial still remember the three points after 24 hours. To test the company's claim, a random sample of 300 persons is selected, of whom 236 can remember the three-point plan after 24 hours. Does this finding substantiate with the advertising company's claim? Use $\alpha = 0.05$.

Step 1

$H_0: \pi = 0.75$

$H_0: \pi \neq 0.75$

Step 2

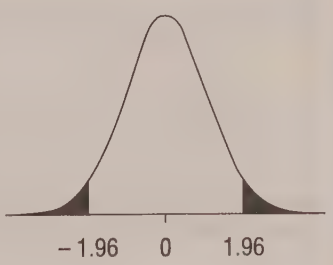
Take $\alpha = 0.05$. Under H_0 the test statistic

$$Z = \frac{P - 0.75}{\sqrt{\frac{0.75(0.25)}{300}}}$$

approximately has a n(0, 1) distribution.

Step 3

Reject H_0 if $z \leq -1.96$ (see accompanying sketch) or if the p-value $= 2\,P(Z \geq |z|) \leq 0.05$.



Step 4

From the above data it follows that $p = \frac{236}{300} = 0.787$ so that

$$Z = \frac{0.787 - 0.75}{\sqrt{\frac{0.75(0.25)}{300}}}$$

$$= 1.48$$

Step 5

Since $-1.96 < z < 1.96$, H_0 is not rejected. Alternatively, the p-value $= 2P(Z \geq |1.48|) = 2(0.5 - 0.43056) = 0.13888$. We can therefore conclude that 75% is a reasonable figure for the percentage of viewers who still remember the three points in the AIDS prevention commercial after 24 hours. Thus the sample does not contradict the claim made by the advertising company.

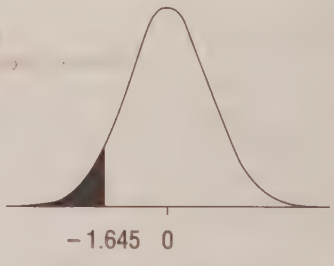
In practice, the estimator of the population parameter in one-tailed hypothesis testing problems often accepts a value that contradicts the direction implied by H_1. In such a case the p-value will be greater than the level of significance, and H_0 will therefore never be rejected. This case is illustrated by the next example.

EXAMPLE 12.11

A management researcher maintains that 60% of companies use incentive schemes at middle-management level. From a representative sample of 138 companies, 86 indicated that they did use incentive schemes. Are the researcher's suspicions confirmed?

$H_0: \pi = 0.60$

$H_1: \pi < 0.60$



We select a significance level of 5%. Under H_0 the test statistic

$$Z = \frac{P - 0.6}{\sqrt{\frac{0.6(0.4)}{138}}}$$

approximately has a n(0, 1) distribution and H_0 is rejected if $z \leq -1.645$.

From the data it follows that $\frac{86}{138} = 0.623$. Due to the fact that the observed sampling ratio is greater than 0.6 (that is, a value in contradiction to the direction of H_1) it follows that H_0 cannot be rejected. The p-value can be calculated as follows:

$$z = \frac{0.623 - 0.6}{\sqrt{\frac{0.6(0.4)}{138}}} = 0.55$$

therefore the p-value $= P(Z \leq z) = P(Z \leq 0.55) = 0.5 + 0.20884 = 0.70884$, which is greater than 0.05.

5.3 The population variance σ^2

In production processes it is extremely important to manufacture articles that meet prescribed homogeneity specifications. It is therefore important to be able to test hypotheses with regard to the population variance σ^2. Example 12.12 illustrates an application of hypothesis testing with regard to σ^2.

EXAMPLE 12.12

The electricity department of a very large city decides to conduct tests on a certain type of mercury-vapour light bulb that, according to the manufacturers, has only a very small variance in its life expectancy. If the manufacturers' claim is valid, the municipality intends installing this type of light bulb in street lighting for all new suburbs in the future. The light bulbs will all be simultaneously replaced after a period of three years, which is slightly less than the mean life expectancy of the bulbs. In this way the municipality intends to eliminate the continual inspection trips and thus to cut long-term costs. The municipality will accept the manufacturer's claim if it can be proved that the standard deviation in the life expectancy of the bulbs is less than 100 hours. The municipality's decision will therefore be based on a choice between

H_0: $\sigma^2 = (100)^2$ and H_1: $\sigma^2 < (100)^2$.

Problems such as the one described in Example 12.12 can be solved by using the procedure summarized in Table 12.5. For this procedure to be valid the random sample $X_1, X_2, \ldots, X_n$ has to be taken from a n(μ, σ^2) population.

Table 12.5 Testing H_0: $\sigma^2 = \sigma_0^2$ on the assumption that the sample observations are taken from a normal distribution

Test statistic	Distribution of test statistic under H_0	Table with critical values
$Y = \dfrac{(n-1)S^2}{\sigma_0^2}$	$\chi^2(n-1)$	Tabel B.4
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
H_1: $\sigma^2 \neq \sigma_0^2$	$y \leq \chi^2_{n-1,\,1-\alpha/2}$ or $y \geq \chi^2_{n-1,\,\alpha/2}$	$\min\{2P(Y \leq y); 2P(Y \geq y)\}$
H_1: $\sigma^2 > \sigma_0^2$	$y \geq \chi^2_{n-1,\,\alpha}$	$P(Y \geq y)$
H_1: $\sigma^2 < \sigma_0^2$	$y \leq \chi^2_{n-1,\,1-\alpha}$	$P(Y \leq y)$

EXAMPLE 12.13

As in various overseas countries, the South African government decides that newly manufactured cars have to meet strict requirements with regard to carbon monoxide (CO) emission. A limit is set at an average of 0.3% CO per unit volume of exhaust fumes. However, since it is possible for a manufacturer to meet this requirement despite the fact that the actual amount of pollution varies considerably from one car to the next, there is an additional stipulation that the standard deviation may not significantly ($\alpha = 0.05$) exceed 0.025% CO particles per unit volume. A random sample of twenty newly manufactured cars of a certain make are tested and the following CO measurements are obtained:

0.292	0.287	0.262	0.274	0.331	0.322	0.325	0.282	0.353	0.299
0.310	0.328	0.315	0.300	0.298	0.229	0.316	0.314	0.283	0.281

Step 1

H_0: $\sigma^2 = (0.025)^2$

H_1: $\sigma^2 > (0.025)^2$

Step 2

$\alpha = 0.05$. Under H_0 the test statistic

$$Y = \frac{19S^2}{(0.025)^2}$$

has a $\chi^2(19)$ distribution.

Step 3

Reject H_0 if $y \geq \chi^2_{19,\,0.05}$ that is, if $y \geq 30.144$.

Step 4

From the data it follows that $s = 0.02792$ so that

$$y = \frac{19 \times (0.02792)^2}{(0.025)^2}$$
$$= 23.698$$

Step 5

Since $y < 30.144$, H_0 is not rejected. The standard deviation of measurements therefore falls within the prescribed specifications. Using a computer program one can show that the p-value $= P(Y \geq 23.698) = 0.20803$. Since the p-value > 0.05, it follows that H_0 cannot be rejected.

6 The two-sample case

In practice, there are often situations in which the parameters of two populations have to be compared. This comparison can be made on the basis of two random samples of size n_1 and n_2. In this section we discuss different tests for the mutual comparison of the population means μ_1 and μ_2, the population ratios π_1 and π_2 and the population standard deviations σ_1 and σ_2.

6.1 Difference between means (large samples)

An important factor in hypothesis testing is to determine whether or not the difference between two observed means $\bar{x}_1$ and $\bar{x}_2$ is due to stochastic variation

and whether or not the difference is big enough to allow the conclusion that the two samples come from populations with different means. Thus, for example, we can try to determine whether or not there is a significant difference in the mean fuel consumption of two car models, or we can compare the mean life expectancy of two types of light bulb. In problems such as these a choice has to be made between H_0: $\mu_1 = \mu_2$, and an appropriate alternative. We first study the case of two independent samples in Table 12.6, in which both n_1 and n_2 are at least 100, with sample variances of S_1^2 and S_2^2 respectively.

The requirement that $n_1 \geq 100$ and $n_2 \geq 100$, is set for the application of the central limit theorem and need not be followed to the letter. If the distributions of the observations can be approximated by symmetrical distributions, sample sizes as small as 30 will also give a satisfactory normal approximation.

Table 12.6 Testing H_0: $\mu_1 - \mu_2 = 0$ on the basis of two independent large samples

Test statistic	**Distribution of test statistic under H_0**	**Table with critical values**
$Z = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$	n_1 and n_2 large: approximately n(0, 1)	Table B.2
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
H_1: $\mu_1 \neq \mu_2$ or H_1: $\mu_1 - \mu_2 \neq 0$	$z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$	$2P(Z \geq \lvert z \rvert)$
H_1: $\mu_1 > \mu_2$ or H_1: $\mu_1 - \mu_2 > 0$	$z \geq z_\alpha$	$P(Z \geq z)$
H_1: $\mu_1 < \mu_2$ or H_1: $\mu_1 - \mu_2 < 0$	$z \leq -z_\alpha$	$P(Z \leq z)$

$z_{0.05} = 1.645$ $z_{0.025} = 1.96$ $z_{0.01} = 2.326$ $z_{0.005} = 2.576$

EXAMPLE 12.14

An insurance company investigates various methods of cutting down on operating costs. One problem is the excessively high monthly telephone account. A preliminary analysis of the account shows that more than half of all outgoing calls are non-official. Since unmarried staff are suspected of being largely responsible, a random sample is drawn from married and unmarried staff without the knowledge of either group. The following table is a summary of the sample information with regard to the total duration of private calls made in one month.

Sample information of non-official calls

Staff	Sample size	Mean	Variance
Single	$n_1 = 65$	$\bar{x}_1 = 5.26$	$s_1^2 = 14.5$
Married	$n_2 = 90$	$\bar{x}_2 = 3.89$	$s_2^2 = 13.7$

Test at a 5% level of significance to see if, on average, private calls by unmarried persons last longer than those by married persons.

Step 1

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 > \mu_2$

where μ_1 and μ_2 are respectively the duration of private calls by unmarried and married persons.

Step 2

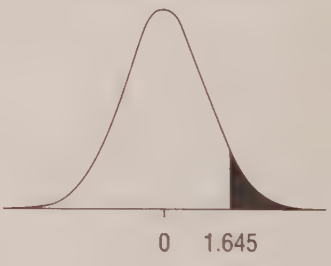
$\alpha = 0.05$. Since n_1 and n_2 are relatively large, the mean duration of the telephone calls should be approximately normally distributed and therefore the test procedure summarized in Table 12.6 can be used. Consequently, under H_0

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{65} + \frac{S_2^2}{90}}}$$

is approximately n(0, 1) distributed.

Step 3

Reject H_0 if $z \geq 1.645$ or if the p-value $= P(Z \geq z) \leq 0.05$.



Step 4

From the data it follows that

$$z = \frac{5.26 - 3.89}{\sqrt{\frac{14.5}{65} + \frac{13.7}{90}}}$$

$$= 2.24$$

Step 5

As $z > 1.645$, H_0 is rejected and we conclude that the mean duration of private calls by unmarried persons is longer than that of calls by married persons. The p-value can be calculated as follows from Table B.2: p-value $= P(Z \geq 2.24) = 0.5 - 0.48745 = 0.01255$.

6.2 Difference between means for small samples, given $\sigma_1^2 = \sigma_2^2$

The procedure explained in Table 12.7 can be used to test whether or not the means are equal for two independent small samples from distributions whose population curves can be approximated by the normal curve given $\sigma_1^2 = \sigma_2^2$.

Table 12.7 Testing H_0: $\mu_1 - \mu_2 = 0$ on the basis of two independent small samples from normal populations with equal variances

Test statistic	Distribution of test statistic under H_0	Table with critical values
$T = \dfrac{\bar{X}_1 - \bar{X}_2}{S\sqrt{1/n_1 + 1/n_2}}$ with $S = \sqrt{\dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$	n_1 and n_2 small: $t(n_1+n_2-2)$	Table B.3
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
H_1: $\mu_1 \neq \mu_2$ or H_1: $\mu_1 - \mu_2 \neq 0$	$t \leq -t_{n_1+n_2-2,\,\alpha/2}$ or $t \geq t_{n_1+n_2-2,\,\alpha/2}$	$2P(T \geq \lvert t \rvert)$
H_1: $\mu_1 > \mu_2$ or H_1: $\mu_1 - \mu_2 > 0$	$t \geq t_{n_1+n_2-2,\,\alpha}$	$P(T \geq t)$
H_1: $\mu_1 < \mu_2$ or H_1: $\mu_1 - \mu_2 < 0$	$t \leq -t_{n_1+n_2-2,\,\alpha}$	$P(T \leq t)$

EXAMPLE 12.15

An agriculturist claims that she has cultivated a new variety of orange that has a significantly higher content of ascorbic acid (vitamin C) than the variety that is the current top seller. The accompanying data shows the ascorbic acid content (mg/100g) for a random sample of fifteen oranges of the new variety and twelve of the current top seller.

Ascorbic content of two varieties of orange

New variety ($n_1 = 15$)			Current top seller ($n_2 = 12$)		
95.1	104.3	100.9	92.5	92.1	90.4
99.2	99.8	99.4	96.4	93.1	90.1
99.3	100.3	103.5	100.8	95.1	96.4
97.5	102.8	99.5	94.5	91.7	93.6
100.3	95.7	100.6			

Let μ_1 and μ_2 indicate the population mean ascorbic acid content of the new variety and of the current top seller. The problem is solved as follows:

Step 1

H_0: $\mu_1 = \mu_2$

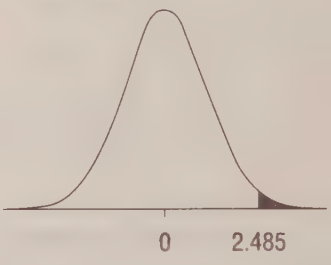
H_1: $\mu_1 > \mu_2$

Step 2

Let $\alpha = 0.01$. The assumption made is that the ascorbic acid content of the two types of orange is normally distributed with equal variances (see Example 12.22) so that under H_0 the test statistic has the t(25) distribution.

Step 3

H_0 is rejected if $t \geq t_{25,\,0.01} = 2.485$
or if the p-value $= P(T \geq t) \leq 0.01$.



Step 4

The data shows that $n_1 = 15$, $\bar{x}_1 = 99.880$ and $s_1 = 2.5389$, $n_2 = 12$, $\bar{x}_2 = 93.892$, and $s_2 = 3.0056$. Thus $s_1^2 = 6.4460$ and $s_2^2 = 9.0336$ so that

$$s = \sqrt{\frac{14 \times 6.4460 + 11 \times 9.0336}{25}}$$

$$= 2.7540 \text{ so that}$$

$$t = \frac{99.880 - 93.892}{2.7540\sqrt{\frac{1}{15} + \frac{1}{12}}}$$

$$= 5.614$$

Step 5

Since $t > 2.845$, H_0 is rejected and we conclude that the new variety of orange has a significantly higher ascorbic acid content than the current top-selling orange. The p-value can be calculated as follows: p-value $= P(T \geq 5.614) = 0.0000038$.

6.3 Difference between means of small samples, given $\sigma_1^2 \neq \sigma_2^2$

In the above example for testing H_0: $\mu_1 = \mu_2$ we simply assumed that $\sigma_1^2 = \sigma_2^2$. To apply this procedure, we first have to confirm that the population variances are in fact equal. The testing procedure for H_0: $\sigma_1^2 = \sigma_2^2$ is explained in Table 12.11. However, when the variances differ significantly, the testing procedure for testing H_0: $\mu_1 = \mu_2$ is as summarized in Table 12.8.

Table 12.8 Testing H_0: $\mu_1 - \mu_2 = 0$ on the basis of two independent small samples from normal populations with unequal variances

Test statistic	Distribution of test under H_0	Table with critical values
$T = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$	t distribution with ν degrees of freedom where $\nu = \dfrac{(S_1^2/n_1 + S_2^2/n_1)^2}{\dfrac{(S_1^2/n_1)^2}{n_1 - 1} + \dfrac{(S_2^2/n_2)^2}{n_2 - 1}}$	Table B.3
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
$H_1: \mu_1 \neq \mu_2$ or $H_1: \mu_1 - \mu_2 \neq 0$	$t \leq -t_{\nu,\alpha/2}$ or $t \geq t_{\nu,\alpha/2}$	$2P(T \geq \lvert t \rvert)$
$H_1: \mu_1 > \mu_2$ or $H_1: \mu_1 - \mu_2 > 0$	$t \geq t_{\nu,\alpha}$	$P(T \geq t)$
$H_1: \mu_1 < \mu_2$ or $H_1: \mu_1 - \mu_2 < 0$	$t \leq -t_{\nu,\alpha}$	$P(T \leq t)$

EXAMPLE 12.16

Two risk factors that have a bearing on the condition of the heart are fitness and cholesterol level. In a research project on this subject the amounts of triglycerides (unsaturated fats) in the blood samples of nine coronary patients and 21 marathon athletes were measured. The observations in millimol per litre are as follows:

Amount of triglycerides (millimol per litre) in blood samples of coronary patients and marathon athletes

Coronary patients $n_1 = 9$			Marathon athletes $n_2 = 21$						
3.80	2.71	1.60	0.86	0.84	1.15	1.12	0.72	1.62	1.23
1.62	1.93	1.32	1.22	1.13	0.98	0.62	0.38	0.86	1.25
1.09	2.28	0.65	0.90	0.56	0.66	0.73	0.73	0.50	0.92

Example 12.21 shows that the variances of the two populations differ significantly. Let μ_1 and μ_2 represent the population mean cholesterol levels of coronary patients and marathon athletes. We want to test the hypothesis that, compared to marathon athletes, coronary patients have a significantly higher population mean triglyceride level at a 1% level of significance. From the above data the following is calculated: $\bar{x}_1 = 1.8889$, $s_1^2 = 0.8917$, $\bar{x}_2 = 0.9038$ and $s_2^2 = 0.0901$. The problem is then solved by applying the procedure given in Table 12.8:

Step 1

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 > \mu_2$

Step 2

$\alpha = 0.01$. The test statistic under H_0 approximately has a t distribution with degrees of freedom equal to

$$\nu = \frac{(0.8917/9 + 0.0901/21)^2}{\frac{(0.8917/9)^2}{8} + \frac{(0.0901/21)^2}{20}}$$

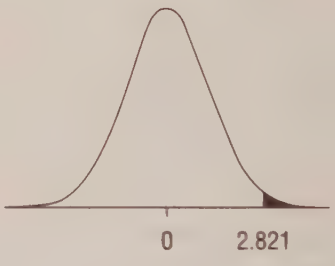
$$= \frac{(0.0991 + 0.0043)^2}{\frac{(0.0991)^2}{8} + \frac{(0.0043)^2}{20}}$$

$$= 8.703$$

$$\approx 9$$

Step 3

Reject H_0 if $t \geq t_{9,\,0.01} = 2.821$ or if the p-value ≤ 0.01.



Step 4

From this data it follows that

$$t = \frac{1.8889 - 0.9038}{\sqrt{\frac{0.8917}{9} + \frac{0.0901}{21}}}$$

$$= 3.064$$

Step 5

Since $t > 2.821$, H_0 is rejected and we conclude that the mean amount of triglycerides for coronary patients is higher than for marathon athletes. By means of a computer program the p-value is calculated as $P(T \geq 3.064) = 0.00701$. If we have assumed that $\sigma_1^2 = \sigma_2^2$, then $t = 4.377$ at $\nu = 28$ degrees of freedom so that the p-value $= 0.000076$ (see Table 12.7). This illustrates the possible effect of an erroneous assumption about the result of a statistical test.

6.4 Difference between means (paired observations)

The test for the difference between means discussed in the above subsections 6.1 to 6.3, is based on the assumption that the two samples on which the test procedures are based, are independent.

Two samples are **independent** if there is no explicable relation between the elements of the one sample and those of the other sample.

However, in practice, there are situations in which we need to compare two population means on the basis of information contained in two dependent samples.

Two samples of equal size are **dependent** when the observations are paired in the sense that each observation from the one sample can be associated with a corresponding observation from the other sample.

When people are selected two at a time on the basis of a common characteristic, for instance IQ, language ability or income, the observations of the characteristic concerned are known as paired observations. Paired observations are also obtained if, for example, the achievements of people are measured before and after completion of a course, or of patients before and after having undergone certain treatment.

The stochastic variable D_i can subsequently be defined as follows in the case of two dependent samples:

D_i = (ith observation from sample 1) – (ith observation from sample 2),
$i = 1, 2, \ldots, n$.

The sample mean and standard deviation of the differences are indicated by $\overline{D}$ and S_D respectively, while d_i, $\overline{d}$ and s_d respectively indicate the observed values of the relevant stochastic variables. The procedure for testing the difference between the two means μ_1 and μ_2 amounts to testing the mean μ_D of the distribution of the differences D_i, and is given in Table 12.9. As in the test procedure in Table 12.3, it is assumed in the case of small samples that the differences D_i are normally distributed. In the case of large samples an approximated test statistic is obtained by applying the central limit theorem. The appropriate test procedure is summarized in Table 12.9.

EXAMPLE 12.17

Over a period of five years it was found that the Boeing 737 aircraft averaged 80 landings with a set of main landing tyres. The South African Airways also established that the landing speed of the aircraft has a significant effect on the life expectancy of the tyres and that tyres supplied by manufacturer A possibly showed wear sooner than those supplied by manufacturer B. One aeroplane was handed over for testing purposes for a period of one year and tyres were fitted to the two sets of landing gear. Tyres from manufacturer A were fitted on the one set of landing gear, and tyres from manufacturer B on the other, so that the tyres from the two manufacturers could be compared under controlled conditions. It is important

Table 12.9 Testing H_0: $\mu_1 - \mu_2 = 0$ on the basis of two dependent samples (paired observations)

Test statistic	Distribution of test statistic under H_0	Table with critical values
$T = \dfrac{\overline{D}}{S_D/\sqrt{n}}$	n small: t($n-1$) n large: approximately n(0, 1)	Table B.3 Table B.2
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
$H_1: \mu_1 \neq \mu_2$ or $H_1: \mu_1 - \mu_2 \neq 0$	$t \leq -t_{n-1,\,\alpha/2}$ or $t \geq t_{n-1,\,\alpha/2}$	$2P(T \geq \lvert t \rvert)$
$H_1: \mu_1 > \mu_2$ or $H_1: \mu_1 - \mu_2 > 0$	$t \geq t_{n-1,\,\alpha}$	$P(T \geq t)$
$H_1: \mu_1 < \mu_2$ or $H_1: \mu_1 - \mu_2 < 0$	$t \leq -t_{n-1,\,\alpha}$	$P(T \leq t)$

to note that the experiment was designed so that, with every landing, all factors that could possibly influence the wear of the tyres were identical for both types of tyre. These factors were, to name but a few, the freight carried by the aeroplane, the landing speed, the temperature of the runway, its wetness and the wind velocity. The data obtained over a period of twelve months is given in the table below:

Number of landings and landing speed on two types of tyre for twelve months

Month	Landing speed (knots)	Number of landings Manufacturer A	Number of landings Manufacturer B	Difference d	d^2
(1)	(2)	(3)	(4)	(5)	(6)
1	112	86	84	2	4
2	114	85	85	0	0
3	116	83	84	–1	1
4	118	79	80	–1	1
5	120	80	81	1	1
6	122	74	76	–2	4
7	124	75	78	–3	9
8	126	67	71	–4	16
9	128	71	72	–1	1
10	130	66	68	–2	4
11	132	67	69	–2	4
12	134	58	64	–6	36
				–19	81

Data obtained from MBL Group PTA00292

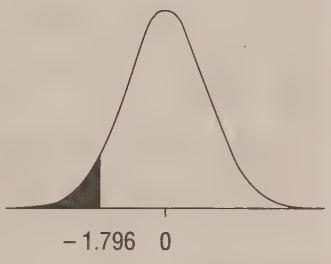
To determine whether or not the tyres from manufacturer A did in fact tend to wear out sooner than the ones from manufacturer B, we have to test whether the mean number of landings with tyres from manufacturer A is smaller than the corresponding figure for tyres from manufacturer B. This is done by calculating d as the differences between columns 3 and 4.

Step 1

$H_0: \mu_1 - \mu_2 = 0$ or $\mu_D = 0$
$H_1: \mu_1 - \mu_2 < 0$ or $\mu_D < 0$

Step 2

$\alpha = 0.05$. Under H_0 the test statistic $\dfrac{\overline{D}}{S_D/\sqrt{12}}$ has a t(11) distribution.



Step 3

Reject H_0 if $t \leq t_{11,0.05} = -1.796$
or if the p-value $= P(T \leq t) \leq 0.05$

Step 4

From the above data it follows that $\overline{d} = \dfrac{-19}{12} = -1.5833$ and that

$$s_d = \sqrt{\frac{1}{n-1}\left\{\sum d^2 - \frac{\left(\sum d\right)^2}{n}\right\}} = \sqrt{\frac{1}{11}\left\{81 - \frac{(-19)^2}{12}\right\}} = 2.1515 \text{ so that}$$

$$t = \frac{-1.5833}{2.1515/\sqrt{12}} = -2.549$$

Step 5

Since $t < -1.796$, H_0 is rejected and we conclude that the mean number of landings for manufacturer A is significantly smaller than for manufacturer B. Using a computer, the p-value can be calculated as follows:

p-value $= P(T \leq -2.549) = 0.01352$.

EXAMPLE 12.18

The time (in minutes) it takes operators to fit a certain part before and after completing a training programme appears in columns 2 and 3 of the table below.

The d values in the fourth column are obtained by subtracting the second observation from the first for each operator. We want to determine whether or not the training programme significantly decreased the mean fitting time, while assuming that the difference in fitting time is normally distributed.

Time required by eight operators before and after training

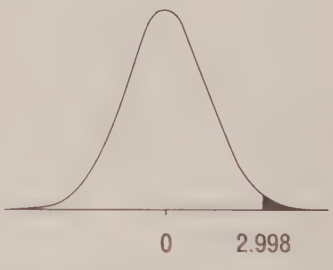
Operator	Before training	After training	d	d^2
(1)	(2)	(3)	(4)	(5)
1	23	17	6	36
2	17	14	3	9
3	16	12	4	16
4	15	13	2	4
5	19	12	7	49
6	21	20	1	1
7	13	14	−1	1
8	20	15	5	25
			27	141

Step 1

H_0: $\mu_1 - \mu_2 = 0$ or H_0: $\mu_D = 0$
H_1: $\mu_1 - \mu_2 > 0$ or H_1: $\mu_D > 0$

Step 2

Let $\alpha = 0.01$. Under H_0 the test statistic $\dfrac{\overline{D}}{S_D/\sqrt{8}}$ has a t(7) distribution.



Step 3

Reject H_0 if $t \geq t_{7,0.01} = 2.998$
or if the p-value $= P(T \geq t) \leq 0.01$.

Step 4

From the above details it follows that $\overline{d} = \dfrac{27}{8} = 3.375$ and that

$$s_d = \sqrt{\frac{1}{7}\left\{141 - \frac{(27)^2}{8}\right\}} = 2.669 \text{ so that}$$

$$t = \frac{3.375}{2.669/\sqrt{8}} = 3.577$$

Step 5

Since $t > 2.998$, H_0 is rejected and we conclude that the training programme significantly decreased the mean fitting time. A computer is used to calculate the p-value as $P(T \geq 3.577) = 0.00451$.

Note

A researcher wants to test the hypothesis

H_0: $\mu_1 - \mu_2 = k$

rather than the usual hypothesis that the two means have the same value. For instance, it might be necessary to test the hypothesis that the mean fuel consumption of one type of car is 2 litres per 100 km higher than that of another. The alternative hypothesis is then one of the following three:

H_1: $\mu_1 - \mu_2 \neq k$ (two-tailed)

H_1: $\mu_1 - \mu_2 > k$ (upper-tailed)

H_1: $\mu_1 - \mu_2 < k$ (lower-tailed)

with k a constant value. In the above example $k = 2$. The test procedure explained in Tables 12.6 to 12.9 is used here, except that the test statistics are adjusted as follows:

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - k}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \quad \text{(see Tables 12.6 and 12.8)}$$

$$T = \frac{\bar{X}_1 - \bar{X}_2 - k}{S\sqrt{1/n_1 + 1/n_2}} \quad \text{with}$$

$$S = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \quad \text{(see Table 12.7)}$$

$$T = \frac{\bar{D} - k}{S_D/\sqrt{n}} \quad \text{(see Table 12.9)}.$$

EXAMPLE 12.19

The braking distance was measured for eleven cars from a speed of 100 km/h under wet conditions and for eleven others under dry conditions. Test whether or not the mean stopping distance under wet conditions is more than 40 metres further than under dry conditions. Let $\alpha = 0.05$. Assume that the variances are equal, and that the stopping distance is normally distributed.

The observations obtained are given in the table:

Stopping distance in metres for 22 cars under different conditions

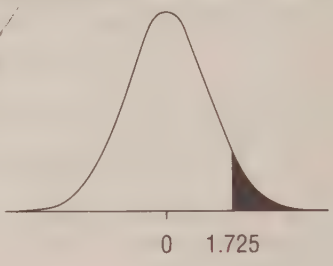
Wet conditions						Dry conditions					
108	138	113	121	126	118	62	59	67	73	69	73
116	106	102	118	140		48	58	66	76	80	

Take μ_1 and μ_2 as the mean stopping distances under wet and dry conditions respectively. The null hypothesis and alternative hypothesis are then formulated as follows:

$$H_0: \mu_1 - \mu_2 = 40 \text{ and } H_1: \mu_1 - \mu_2 > 40.$$

The appropriate test procedure is the one used for two independent small samples with equal variances, as summarized in Table 12.8, but the test statistic is modified as in the note above. Under H_0 the test statistic

$$T = \frac{X_1 - X_2 - 40}{S\sqrt{\frac{1}{11} + \frac{1}{11}}} \text{ approximately has a t(20) distribution.}$$



We therefore reject H_0 if $t = t_{20,0.05} \geq 1.725$
or if the p-value = $P(T \geq t) \leq 0.05$.
The means and sample variances are the following:

$$\bar{x}_1 = 118.727,$$
$$s_1^2 = 148.018,$$
$$\bar{x}_2 = 66.455 \text{ and}$$
$$s_2^2 = 85.473$$
$$s = \sqrt{\frac{10 \times 148.018 + 10 \times 85.473}{20}}$$
$$= 10.805$$

Therefore

$$t = \frac{118.727 - 66.455 - 40}{10.805\sqrt{\frac{1}{11} + \frac{1}{11}}}$$
$$= 2.664$$

Since $t > 1.725$, H_0 is rejected. The stopping distance from a speed of 100 km/h under wet conditions is more than 40 metres further than under dry conditions. The p-value = $P(T \geq 2.664) = 0.00745$. The reader can now test whether or not the assumptions of equal variances (see Table 12.11) and normality (see Chapter 10, Section 7) are justifiable.

6.5 Equality of proportions

Questions such as the following are often important: Is the proportion of adults in Cape Town who eat maize porridge the same as the corresponding proportion in Pretoria? Is the percentage of unmarried women who are subscribing members of a particular political party the same as that of married women? Is the proportion of the incidence of brucellosis infection among cows the same in two different regions? Is the incidence of lung cancer among non-smokers the same as that among smokers? These questions can be answered by testing

$$H_0: \pi_1 = \pi_2$$

against a suitable alternative. Table 12.10 provides a summary of the test procedure based on two independent samples of sizes n_1 and n_2. It must also be assumed that the requirements for the normal approximation of the binomial distribution (see Chapter 10, Section 8) have been met.

Table 12.10 Testing H_0: $\pi_1 - \pi_2 = 0$ under the assumption that the difference in sampling proportions $P_1 - P_2$ is approximately normally distributed

Test statistic	Distribution of test statistic under H_0	Table with critical values
$Z = \dfrac{P_1 - P_2}{\sqrt{P(1-P)(\frac{1}{n_1} + \frac{1}{n_2})}}$ with $P = \dfrac{X_1 + X_2}{n_1 + n_2}$	n_1 and n_2 large: approximately n(0, 1)	Table B.2
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
$H_1: \pi_1 \neq \pi_2$ or $H_1: \pi_1 - \pi_2 \neq 0$	$z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$	$2P(Z \geq \lvert z \rvert)$
$H_1: \pi_1 > \pi_2$ or $H_1: \pi_1 - \pi_2 > 0$	$z \geq z_\alpha$	$P(Z \geq z)$
$H_1: \pi_1 < \pi_2$ or $H_1: \pi_1 - \pi_2 < 0$	$z \leq -z_\alpha$	$P(Z \leq z)$

$z_{0.05} = 1.645 \quad z_{0.025} = 1.96 \quad z_{0.01} = 2.326 \quad z_{0.005} = 2.576$

EXAMPLE 12.20

One of the most cost-effective methods of conducting market research is to use questionnaires that are sent out by post. A significant disadvantage attached to this method of collecting data is that usually a large percentage of respondents will not complete the questionnaires. In many investigations, typically only 30% or less of the questionnaires are returned. Usually the section of the sample that does respond cannot be regarded as being representative of the population, and sampling bias occurs. At a specific marketing research institution a new manager is appointed and he is convinced that, to a large extent, the problem of bias can be eliminated by redesigning the questionnaires. He instructs the production section to design an attractive questionnaire in colour, and to illustrate some of the questions with photographs. To test the effect of the new questionnaire, 750 old questionnaires and 750 new ones are posted. The number of questionnaires returned is given in the table below.

Summary of response to two types of questionnaire

Type of questionnaire	Response	No response	Total
Standard questionnaire	245	505	750
Attractive questionnaire	298	452	750
Total	543	957	1 500

Test at the 1% level of significance, that the response to the standard questionnaire is lower than the response to the attractive questionnaire. Let π_1 and π_2 indicate the population ratios of responses to the standard questionnaire and to the attractive questionnaire respectively.

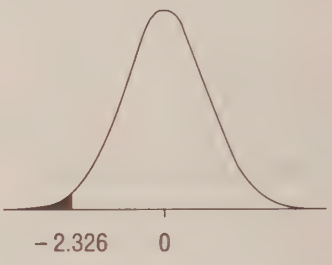
Step 1

H_0: $\pi_1 = \pi_2$

H_1: $\pi_1 < \pi_2$

Step 2

Let $\alpha = 0.01$. Under H_0 the test statistic approximately has a n(0, 1) distribution.



Step 3

Reject H_0 if $z \leq z_{0.01} = -2.326$ or if the p-value ≤ 0.01.

Step 4

From the data we calculate that

$$p_1 = \frac{245}{750} = 0.327,$$

$$p_2 = \frac{298}{750} = 0.397, \text{ and}$$

$$p = \frac{543}{1500} = 0.362 \text{ so that}$$

$$z = \frac{0.327 - 0.397}{\sqrt{0.362(0.638)\left(\frac{1}{750} + \frac{1}{750}\right)}} = -2.82$$

Step 5

Since $z < -2.326$, H_0 is rejected. There is a very significant increase in response to the attractive questionnaire compared to the response to the standard questionnaire. The p-value is

$$P(Z \leq -2.82) = 0.5 - 0.49760 = 0.00240$$

6.6 Difference in variances

Consider the two independent samples of sizes n_1 and n_2. If the underlying distributions are normal, the procedures explained in Table 12.18 can be used to test

$$H_0: \sigma_1^2 = \sigma_2^2$$

against a suitable alternative. This type of testing problem is encountered when a firm has to choose, for example, between two types of parts that have the

same mean life expectancy. Usually, the firm will select the part with the smallest variance, provided other factors such as the purchase price do not play a role. This null hypothesis often has to be tested before the null hypothesis of equal means H_0: $\mu_1 = \mu_2$ is tested, particularly in the case of small samples (see Tables 12.7 and 12.8). Note that f in Table 12.11 indicates an observed value of the F distribution.

Table 12.11 Testing H_0: $\sigma_1^2 = \sigma_2^2$ on the basis of two independent samples of sizes n_1 and n_2 from two normal populations

Test statistic	Distribution of test statistic under H_0	Table with critical values
$F = \dfrac{S_1^2}{S_2^2}$	$F(n_1-1, n_2-1)$	Table B.5
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
$H_1: \sigma_1^2 \neq \sigma_2^2$ or $H_1: \dfrac{\sigma_1^2}{\sigma_2^2} \neq 1$	$f \leq F_{n_1-1, n_2-1; 1-\alpha/2}$, that is $f \leq \dfrac{1}{F_{n_2-1, n_1-1; \alpha/2}}$ or $f \geq F_{n_1-1, n_2-1; \alpha/2}$	$\min\{2P(F \leq f); 2P(F \geq f)\}$
$H_1: \sigma_1^2 > \sigma_2^2$ or $H_1: \dfrac{\sigma_1^2}{\sigma_2^2} > 1$	$f \geq F_{n_1-1, n_2-1; \alpha}$	$P(F \geq f)$
$H_1: \sigma_1^2 < \sigma_2^2$ or $H_1: \dfrac{\sigma_1^2}{\sigma_2^2} < 1$	$f \leq F_{n_1-1, n_2-1; 1-\alpha}$, that is $f \leq \dfrac{1}{F_{n_2-1, n_1-1; \alpha}}$	$P(F \leq f)$

EXAMPLE 12.21

In Example 12.16 we assumed that the variances in the triglyceride levels for coronary patients and marathon athletes were unequal. We now illustrate how this assumption is tested. For the testing procedure we assume that the two underlying populations of triglyceride levels are normally distributed (see normal probability plot, Chapter 10, Section 7). The relevant data is as follows:

$n_1 = 9$ with $s_1^2 = 0.8917$ and $n_2 = 21$ with $s_2^2 = 0.0901$.

The relevant hypotheses are

H_0: $\sigma_1^2 = \sigma_2^2$

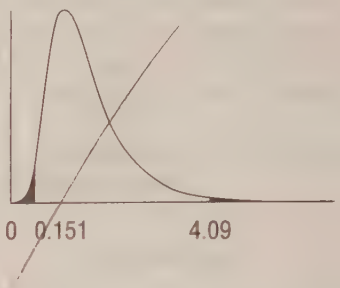
H_1: $\sigma_1^2 \neq \sigma_2^2$

Take the level of significance as $\alpha = 0.01$. The test statistic $F = \frac{S_1^2}{S_2^2}$ has an F(8, 20) distribution under H_0 and we reject H_0 if (Table B.5)

$$f \leq F_{8,\,20;0.995} \text{ or } f \geq F_{8,\,20;0.005}$$

that is if

$$f \leq \frac{1}{F_{20,\,8;\,0.005}} = 0.151 \text{ or } f \geq 4.09$$



From the data it follows that

$$f = \frac{s_1^2}{s_2^2} = \frac{0.8917}{0.0901} = 9.90,$$

so H_0 is rejected. The assumption of unequal variances is therefore justified. The p-value = 0.000035.

EXAMPLE 12.22

In Example 12.15 we assumed that the ascorbic acid contents of the two types of orange were normally distributed with equal variances. We can now test whether the assumption of equal variances is justified with $\alpha = 0.05$. The data is as follows: $n_1 = 15$, with $s_1^2 = 6.4460$ and $n_2 = 12$ with $s_2^2 = 9.0336$.

The relevant hypotheses are

H_0: $\sigma_1^2 = \sigma_2^2$

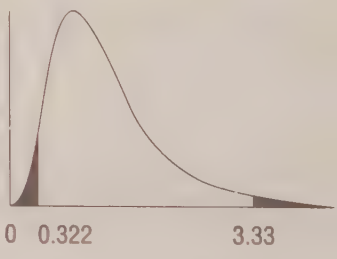
H_1: $\sigma_1^2 \neq \sigma_2^2$

Take the level of significance as $\alpha = 0.05$. The test statistic $F = \frac{S_1^2}{S_2^2}$ has an F(14, 11) distribution under H_0 and we reject H_0 if (Table B.5)

$$f \leq F_{14,\,11;\,0.975} \text{ or } f \geq F_{14,\,11;\,0.025},$$

that is if

$$f \leq \frac{1}{F_{11,\,14;\,0.025}} = \frac{1}{3.1} = 0.322 \text{ or } f \geq 3.33$$



From the data it follows that

$$f = \frac{s_1^2}{s_2^2} = \frac{6.4460}{9.0336} = 0.71$$

so that H_0 is not rejected. The assumption of unequal variances is therefore justified. The p-value = 0.53872.

7 The power of a test

In Section 3 the probability of making a type II error is expressed as

$$\beta = P(\text{type II error})$$
$$= P(H_0 \text{ accepted} \mid H_1).$$

Because H_1 is usually a compound hypothesis, β is a function of the values that can be assumed by the population parameter. The important concepts in this section can best be explained with the aid of an example.

Consider upper-tailed testing of the proportion of successes, π, for a binomial experiment with

$H_0: \pi = \pi_0$

$H_1: \pi > \pi_0$

The alternative H_1 includes an interval of π values, namely $\{\pi_1 \mid \pi_0 < \pi_1 < 1\}$. The probability of a type II error can therefore be expressed as follows for this interval:

$$\beta(\pi_1) = P(Z \leq z_\alpha \mid \pi_1)$$

$$= P\left(\frac{P - \pi_0}{\sigma_{P0}} \leq z_\alpha \mid \pi_1\right)$$

where (see Chapter 10, Section 10) $\sigma_{P0} = \sqrt{\pi_0(1 - \pi_0)/n}$ is the standard deviation of the sample ratio of successes, P, if $\pi = \pi_0$. Note that if π_1 is the actual proportion of successes, $\frac{P - \pi_0}{\sigma_{P0}}$ does not have a n(0, 1) distribution. In this case $Z = \frac{P - \pi_1}{\sigma_{P1}}$ has a n(0, 1) distribution with $\sigma_{P1} = \sqrt{\pi_1(1 - \pi_1)/n}$ and $\beta(\pi_1)$ can therefore be calculated as follows:

$$\beta(\pi_1) = P(P \leq z_\alpha \sigma_{P0} + \pi_0 \mid \pi_1)$$

$$= P(P - \pi_1 \leq z_\alpha \sigma_{P0} + \pi_0 - \pi_1)$$

$$= P\left(\frac{P - \pi_1}{\sigma_{P1}} \leq \frac{z_\alpha \sigma_{P0} + \pi_0 - \pi_1}{\sigma_{P1}}\right)$$

$$= P(Z \leq b) \text{ with } b = \frac{\pi_0 - \pi_1 + z_\alpha \sigma_{P0}}{\sigma_{P1}}$$

The power of this upper-tailed test is now given by

$$P_R(\pi_1) = 1 - \beta(\pi_1) = 1 - P(Z \leq b),$$

where Z is n(0, 1) distributed and b is defined above. The notation P_R indicates the upper-tailed (or right-tailed) nature of the test.

The power of a test is therefore the probability of rejecting H_0 if H_1 is true, that is, the probability of making the right decision.

The **power** of a test for $H_0: \theta = \theta_0$ against the alternative $H_1: \theta \in A_1$, with A_1 a specified domain for θ, is defined by

$$P(\theta_1) = 1 - \beta(\theta_1)$$

for all θ_1 in the domain A_1 and with $\beta(\theta_1)$ the probability of a type II error.

In Section 3 a graphic representation of β, the probability of a type II error, is given (the so-called operating characteristic curve – see Figure 12.1). Consider the following example in which the power of a test is calculated and represented graphically.

EXAMPLE 12.23

A cancer research association claims that almost 50% of all cancer sufferers can be cured provided the right treatment is given in time. An independent research team that has access to the medical history of a large number of cancer patients, decides to examine this claim by testing the following hypothesis at a 5% level of significance:

H_0: $\pi = 0.45$ against H_1: $\pi > 0.45$

From a random sample of the particular target population of cancer patients the research team discovers that, out of 120 patients, 58 currently show no sign of the disease. Determine and draw the power function $P_R(\pi_1)$.

For this problem the following information is available:

$$n = 120,\ \pi_0 = 0.45,\ z_{0.05} = 1.645,\ A_1 = \{\pi_1 \mid 0.45 < \pi_1 < 1\}$$

From the definition of $P_R(\pi_1)$ above we find that

$$\sigma_{P0} = \sqrt{\frac{0.45 \times 0.55}{120}} = 0.04541$$

so that

$$P_R(\pi_1) = 1 - \beta(\pi_1) = 1 - P(Z \leq b) \text{ with } b = \frac{\pi_0 - \pi_1 + 1.645\,\sigma_{P0}}{\sigma_{P1}}$$

so that

$$P_R(\pi_1) = 1 - P\left(Z \leq \frac{0.45 - \pi_1 + 1.645 \times 0.04541}{\sigma_{P1}}\right) \text{ with } \sigma_{P1} = \sqrt{\frac{\pi_1(1 - \pi_1)}{120}}$$

If $\pi_1 = 0.46$ it follows that, for instance, $\sigma_{P1} = 0.04550$.

The following table contains the values of π_1 and $P_R(\pi_1)$ for various values of π_1 from A_1.

Calculation of the power of the test
H_0: $\pi = 0.45$ against H_1: $\pi > 0.45$

π_1	$P_R(\pi_1)$
0.45	0.04999
0.46	0.07748
0.47	0.11493
0.48	0.16348
0.49	0.22346
0.50	0.29415
0.52	0.45890
0.54	0.63161
0.56	0.78197
0.58	0.89013
0.60	0.95387
0.65	0.99800
0.70	1.0

The power function is represented in Figure 12.4.

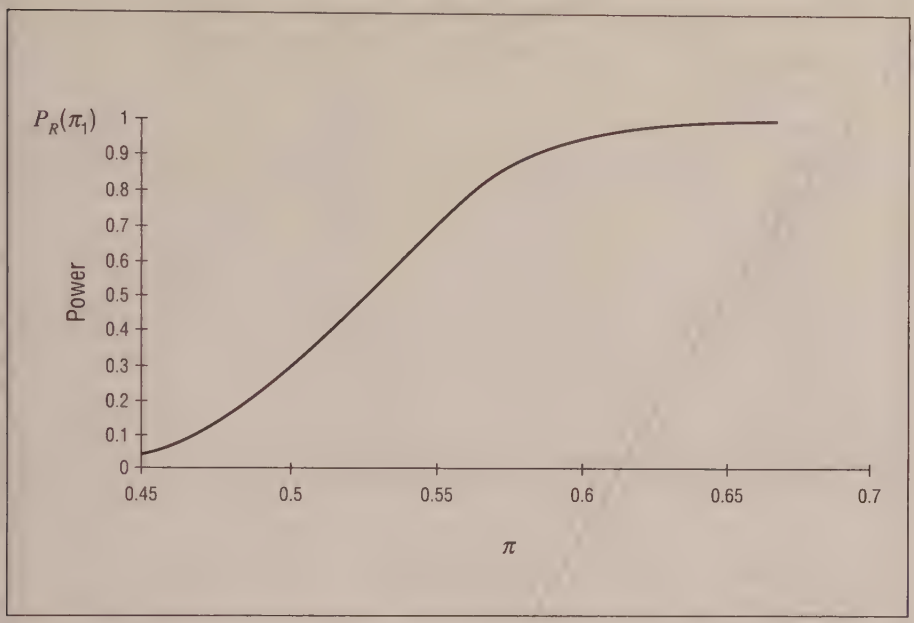


Figure 12.4 The power function for testing H_0: $\pi = 0.45$ against H_1: $\pi > 0.45$

In cases of lower-tailed and two-tailed testing the power is established similarly. In the case of lower-tailed testing, the power of the test for H_0: $\pi = \pi_0$ against H_1: $\pi < \pi_0$ is given by:

$P_L(\pi_1) = 1 - P(Z \geq a)$, where $a = \dfrac{\pi_0 - \pi_1 - z_\alpha \sigma_{P0}}{\sigma_{P1}}$ and the notation P_L denotes the lower-tailed nature of the test.

In the case of two-tailed testing for H_0: $\pi = \pi_0$ against H_1: $\pi = \pi_0$ the power is given by:

$$P_T(\pi_1) = 1 - P(a \leq Z \leq b), \text{ where } a = \frac{\pi_0 - \pi_1 - z_{\alpha/2}\sigma_{P0}}{\sigma_{P1}}$$

$b = \dfrac{\pi_0 - \pi_1 + z_{\alpha/2}\sigma_{P0}}{\sigma_{P1}}$ and the notation P_T denotes a two-tailed alternative.

In Figures 12.5 and 12.6 examples are given of the graphic representation of the power of the test in the lower-tailed and two-tailed cases respectively.

The power of a test cannot always be determined by a simple procedure. However, Table 12.2 contains a summary of the power expressions for a number of one-sample problems. The application of the table is illustrated by the following example for testing

H_0: $\mu = \mu_0$ against H_1: $\mu \neq \mu_0$.

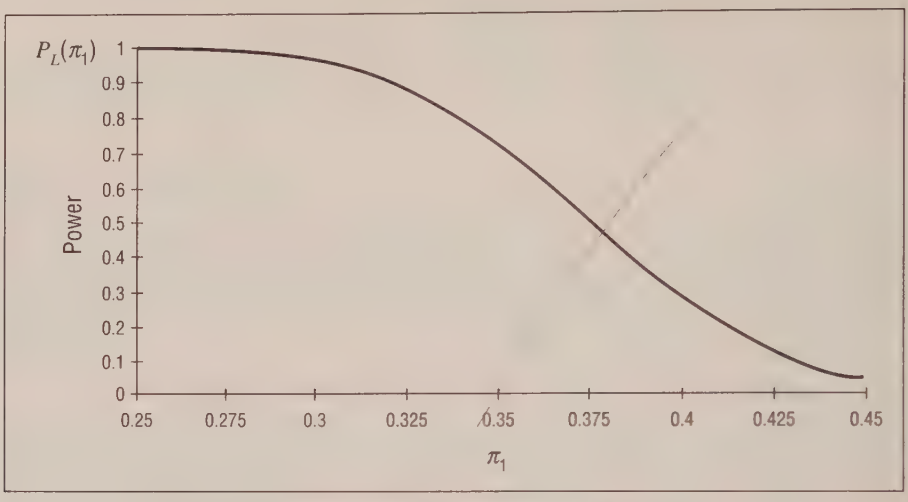


Figure 12.5 The power of the test for H_0: $\pi = \pi_0$ against H_1: $\pi < \pi_0$

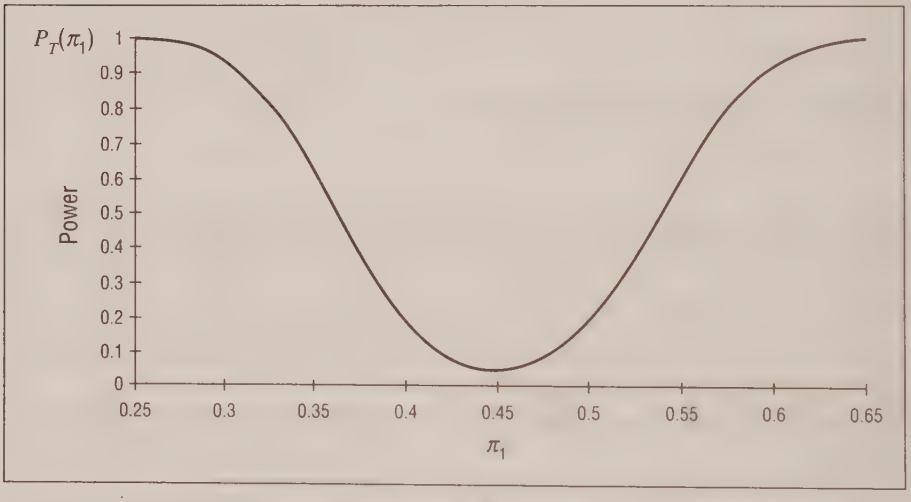


Figure 12.6 The power of the test for H_0: $\pi = \pi_0$ against H_1: $\pi \neq \pi_0$

EXAMPLE 12.24

According to specifications, ball bearings designed for use in tumble dryers must have a cross-section of 50 $\pm$ 0.05 mm. The population standard deviation is unknown and must therefore be estimated. A random sample of 120 ball bearings yields a standard deviation of $s = 0.083$ mm. The company producing the tumble dryers requires the power of the test for H_0: $\mu = 50$ against H_1: $\mu \neq 50$ if $\alpha = 0.05$ for values of μ in the interval 49.95 to 50.05 mm.

Table 12.12 Power functions for the parameters μ, π and σ^2 at an α level of significance

Null hypothesis	Distribution assumption	Alternative hypothesis	Power	Explanation of symbols	Remarks
$H_0: \mu = \mu_0$	n large enough for application of the central limit theorem	$H_1: \mu \neq \mu_0$ $H_1: \mu > \mu_0$ $H_1: \mu < \mu_0$	$P_T(\mu_1) = 1 - P(a \leq Z \leq b)$ $P_R(\mu_1) = 1 - P(Z \leq b)$ $P_L(\mu_1) = 1 - P(Z \geq a)$	$a = c - z_{\alpha/2}$; $b = c + z_{\alpha/2}$ $b = c + z_\alpha$ $a = c - z_\alpha$	Z approximately has a n(0, 1) distribution. Estimate σ by s $c = \dfrac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}$
$H_0: \mu = \mu_0$	$X_1, X_2, \ldots, X_n$ a random sample from a $n(\mu, \sigma^2)$ population, σ^2 unknown	$H_1: \mu \neq \mu_0$ $H_1: \mu > \mu_0$ $H_1: \mu < \mu_0$	$P_T(\mu_1) \approx 1 - P(a \leq Z \leq b)$ $P_R(\mu_1) \approx 1 - P(Z \leq b)$ $P_L(\mu_1) \approx 1 - P(Z \geq a)$	$a = c - t_{n-1;\,\alpha/2}$; $b = c + t_{n-1;\,\alpha/2}$ $b = c + t_{n-1;\,\alpha}$ $a = c - t_{n-1;\,\alpha}$	Z approximately has a n(0, 1) distribution. Estimate σ by s $c = \dfrac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}$
$H_0: \pi = \pi_0$	n large	$H_1: \pi \neq \pi_0$ $H_1: \pi > \pi_0$ $H_1: \pi < \pi_0$	$P_T(\pi_1) = 1 - P(a \leq Z \leq b)$ $P_R(\pi_1) = 1 - P(Z \leq b)$ $P_L(\pi_1) = 1 - P(Z \geq a)$	$a = c - z_{\alpha/2} \times d$; $b = c + z_{\alpha/2} \times d$ $b = c + z_\alpha \times d$ $a = c - z_\alpha \times d$	Z approximately has a n(0, 1) distribution $\sigma_{P0} = \sqrt{\pi_0(1 - \pi_0)/n}$ $\sigma_{P1} = \sqrt{\pi_1(1 - \pi_1)/n}$ $c = \dfrac{\pi_0 - \pi_1}{\sigma_{P1}}$; $d = \dfrac{\sigma_{P0}}{\sigma_{P1}}$
$H_0: \sigma^2 = \sigma_0^2$	$X_1, X_1, \ldots, X_n$ a random sample from a $n(\mu, \sigma^2)$ population	$H_1: \sigma^2 \neq \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$	$P_T(\sigma_1^2) = 1 - P(a \leq Y \leq b)$ $P_R(\sigma_1^2) = 1 - P(Y \leq b)$ $P_L(\sigma_1^2) = 1 - P(Y \geq a)$	$a = c \times \chi^2_{n-1,\,1-\alpha/2}$; $b = c \times \chi^2_{n-1,\,\alpha/2}$ $b = c \times \chi^2_{n-1,\,\alpha}$ $a = c \times \chi^2_{n-1,\,1-\alpha}$	Y has a χ^2 $(n-1)$ distribution $c = \left(\dfrac{\sigma_0}{\sigma_1}\right)^2$

The power function is calculated as follows:

Using Table 12.12 it follows that

$P_T(49.95) = 1 - P(a \leq Z \leq b)$, where $a = \dfrac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} - z_{0.025}$ and $b = \dfrac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{0.025}$ with $\mu_1 = 49.95$.

Since we have estimated σ as 0.083, it follows that

$$a = \frac{50 - 49.95}{0.083/\sqrt{120}} - 1.96 = 4.64 \text{ and}$$

$$b = \frac{50 - 49.95}{0.083/\sqrt{120}} + 1.96 = 8.56$$

Therefore $P_T(49.95) = 1 - P(4.64 \leq Z \leq 8.56) = 1.0$.

Similarly the $P_T(\mu_1)$ values are calculated for $\mu_1 = 49.95, 49.97$ up to and including 50.05. These values appear in the following table. The last column of the same table contains the power. The power of the test for $n = 120$ and $n = 300$ is represented graphically in Figure 12.7.

Power for testing H_0: $\mu = \mu_0$ against H_1: $\mu \neq \mu_0$ for $n = 120$ and for different values of μ_1

μ_1	$a = \dfrac{\mu_0 - \mu_1}{s/\sqrt{n}} - z_{0.025}$	$b = \dfrac{\mu_0 - \mu_1}{s/\sqrt{n}} + z_{0.025}$	$P_T(\mu_1)$
49.95	4.64	8.56	1.00000
49.96	3.32	7.24	0.99955
49.97	2.00	5.92	0.97722
49.98	0.68	4.60	0.75163
49.99	–0.64	3.28	0.26155
50.00	–1.96	1.96	0.05000
50.01	–3.28	0.64	0.26155
50.02	–4.60	–0.68	0.75163
50.03	–5.92	–2.00	0.97722
50.04	–7.24	–3.32	0.99955
50.05	–8.56	–4.64	1.00000

The manufacturer of the ball bearings can draw the following conclusion from this table and from Figure 12.7: in future sampling and testing this procedure has the ability to indicate with a high probability whether or not the manufactured ball bearings meet the specifications, provided n is at least 120.

Notes

Since it is sometimes cumbersome or difficult to establish the power of a test, its practical importance is often underestimated and even ignored. Before a project is undertaken, its cost, in terms of the amount of information required (sample size), should be considered in the light of

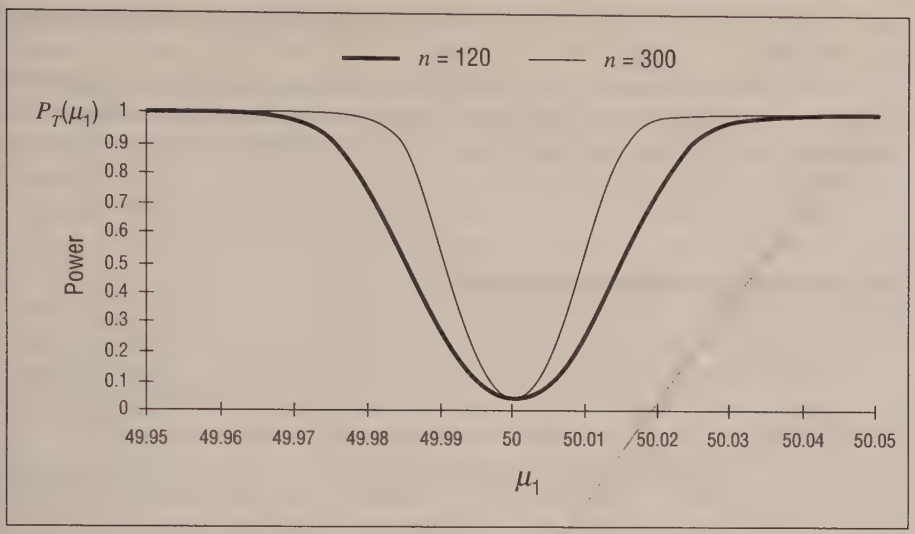


Figure 12.7 The power for testing H_0: $\mu = 50$ against H_1: $\mu \neq 50$ for $n = 120$ and $n = 300$

(a) the risks that will be acceptable (α and β); and

(b) the size of the effects or differences that are important (H_0 and H_1). The relations between these aspects of a project can be studied scientifically only if the power is determined, at the very least, by approximation.

8 Concluding remarks

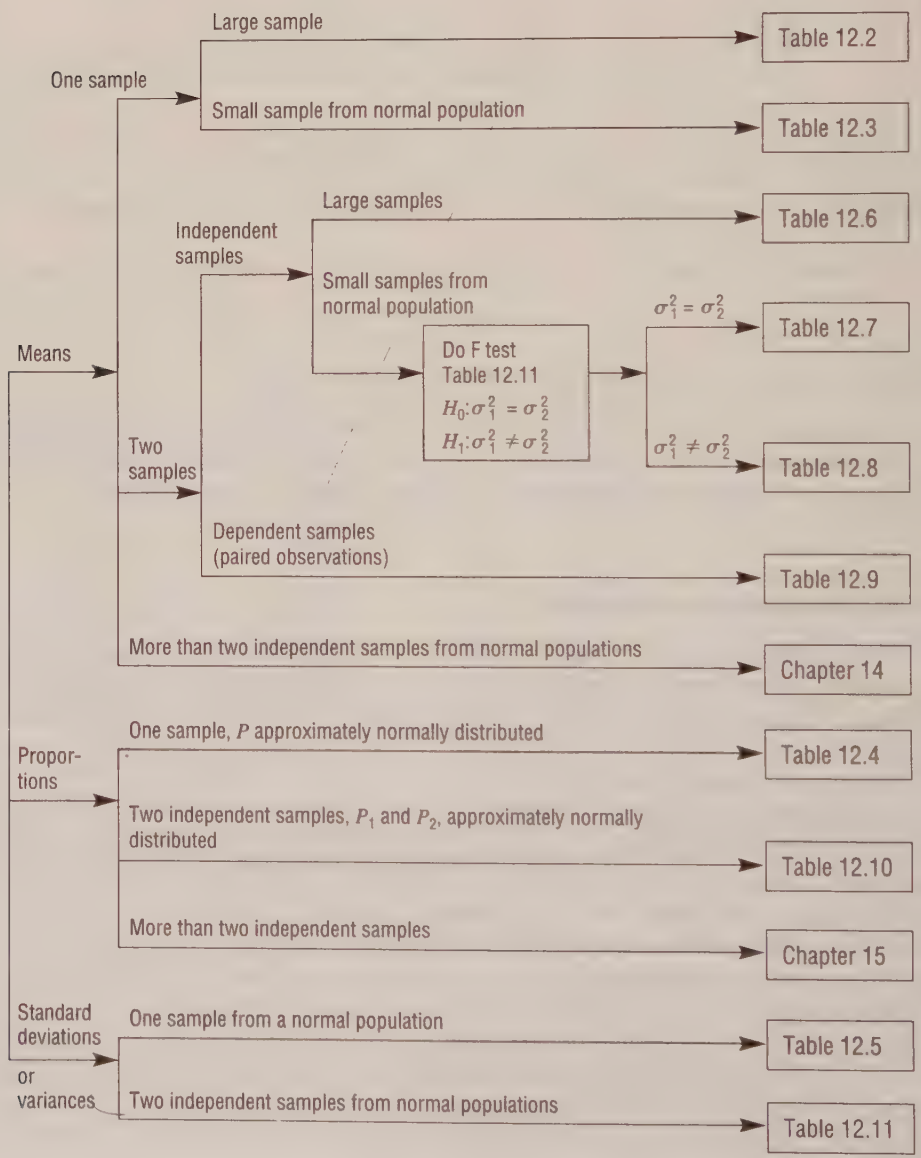
In the preceding sections we have discussed a variety of techniques. For the reader's convenience a guide to hypothesis testing is given in a diagram. A choice between H_0 and H_1 can often be made on the basis of the sample information. However, the procedures we have studied do have limitations, which will be briefly discussed.

Firstly, the power of each procedure is a function of the particular sample size(s). If the sample is large enough, H_0 will always be rejected if the H_0 value of the relevant population parameter and its sample estimate differ, even if this difference is very small (see, for example, Figure 12.7, from which it is clear that the power can be increased by increasing the sample size n). Suppose for instance that a spokesperson for a municipality claims that the mean site rates paid by home owners in a certain suburb during 1991 amounted to R215 per month. If a random sample of 10 000 home owners yields an arithmetic mean of $\bar{x} = \text{R}214.44$ and H_0 is rejected on the basis of this data, we can hardly say that the spokesperson's claim was invalid. For all practical purposes the claim is valid, but the decision maker has to use common sense when interpreting

such a result. The rejection of H_0 in the example might well have only theoretical value.

It is known that the t tests used to test hypotheses regarding population means are relatively **robust** with respect to deviations from the assumption of normality. This implies that there will be very little difference between the

Guide to hypothesis testing



p-values that are calculated and the real (but unknown) values in situations in which the observed data is not normally distributed even in cases where the form of the distribution differs considerably from the normal distribution.

Tests with regard to population spread, for example $\sigma^2 = \sigma_0^2$ and H_0: $\sigma_1^2 = \sigma_2^2$, are particularly sensitive to deviations in the assumption of normality. This means that if the observations are not normally distributed, the conclusions and calculated p-values might well be completely inaccurate (see Chapter 10, Example 10.25). Consequently we recommend that the χ^2 and F tests for spread be used circumspectly and not routinely. If a p-value of, for instance, $p = 0.0001$ is found, it would be relatively safe to conclude that there are significant differences in spread, even if the underlying distributions are not normal. In this example the real (but unknown) risk of making a type I error would in any case be negligible.

The last shortcoming of the procedures studied so far is the assumption that the samples are taken from normal populations. In practice, this assumption is often unfounded and we are forced to turn to alternative testing procedures. In Chapter 16 we consider a number of so-called **distribution-free** techniques where the assumption of normality is not required. There are also procedures that can be used to test whether a sample is taken from a particular population. Such a procedure is, for example, the χ^2 goodness-of-fit test discussed in Chapter 15.

Exercises

1 The manufacturer of rechargeable electric shavers maintains that the mean working time of his product between charges is four hours. A random sample of fourteen shavers gives the following results (in minutes):

260	264	247	255	239	275	257
258	246	286	253	245	262	259

Test the manufacturer's allegation at a 5% level of significance (assume normality).

2 Test the claim that the mean selling price of residences in Pretoria in the period October to December 1991 was R140 000. Assume that the selling prices were normally distributed and use the last twenty figures for the Pretoria residences (Appendix A, Data Set 1, column 3, houses 415 to 434). Take $\alpha = 0.01$.

3 The mean braking time used to accelerate a certain make of 1 800 cm^3 car from 100 km/h to a dead stop is 3.95 seconds. A new braking system designed by the engineering section of the relevant company will allegedly decrease the braking time. In order to test the new braking system, fourteen cars are selected at random and equipped with the system. The braking

times of fourteen crash stops from 100 km/h are as follows (data in seconds):

3.96	3.44	3.36	3.27	3.84	3.49	3.57
3.26	3.84	3.66	4.01	3.36	3.65	3.19

What is your conclusion? ($\alpha = 0.05$) Assume that the stopping times are normally distributed.

4 The diastolic blood pressure of a healthy person can be regarded as a stochastic variable that is normally distributed with $\mu = 85$ mmHg. A certain disease is a suspected cause of high blood pressure. Can this suspicion be confirmed statistically if, for 40 people afflicted with the disease it is found that $\bar{x} = 88.25$ and $s = 5.68$? Let $\alpha = 0.05$. Calculate the p-value.

5 Assume that the number of bacteria per unit volume of water is normally distributed and that a mean count of 20 000 bacteria per unit volume is still safe. The health department of a certain city suspects that this safety limit is being exceeded in one of its dams. A random sample of twelve water samples is drawn and the number of bacteria in each sample is counted. The results are as follows:

24 632	22 568	23 656	21 543	20 346	21 345
19 434	20 995	20 541	21 433	21 854	22 337

Does the health department have cause for concern? Test at the 1% level of significance.

6 It is known that the mean vocabulary of children at a specific age is 1 040 words. An educational researcher has developed a special teaching programme to extend the vocabulary of children in this age group. He has tested his programme on a random sample of 120 children and found that, at the end of the testing period, $\bar{x} = 1\,086$ and $s = 145.3$. Does $\bar{x}$ differ significantly from 1 040 on the 1% level of significance? Calculate the p-value as well.

7 A manufacturer of penicillin claims that less than 12% of all people are allergic to its product. To test the claim, a random sample of 90 persons is drawn. They are treated with penicillin and 10 display an allergic reaction. Does this support the manufacturer's claim? ($\alpha = 0.05$) Calculate the p-value for this test as well.

8 A soft drink company considers converting from the 500 ml glass containers to 500 ml cardboard containers. Unless significantly more than 75% of the consumers prefer the cardboard containers, the change will not be introduced. An opinion poll among 260 consumers shows that 196 are

in favour of the proposed change. What would your recommendation be? Let $\alpha = 0.01$ and calculate the p-value for this test.

9 A manufacturer of car parts claims that less than 10% of all parts supplied to an assembly plant fails to comply with the given specifications. A test of 235 parts produces nineteen defective parts. Test at a 5% level of significance whether the manufacturer's claim is acceptable. Calculate the p-value for this test as well.

10 Seventy per cent of all residences sold in the Johannesburg area are said to be old houses. Use the information on the residences with numbers 241 to 300 (Appendix A, Data Set 1, column 7) to test this allegation. Let $\alpha = 0.01$ and calculate the p-value for this test as well.

11 A random sample of 250 people in Cape Town reveals that 135 watch the news on television regularly. Test whether the population ratio of news viewers is 0.70. Let $\alpha = 0.05$ and calculate the p-value for this test.

12 In an investigation of the life expectancy of a certain type of mercury vapour discharge bulb, the electricity department of a municipality took a random sample of 27 bulbs for testing. It was found that $s = 93.57$ hours. Does this result support the manufacturer's claim that the standard deviation is less than 100 hours? (Assume that the life expectancy is normally distributed and use $\alpha = 0.05$.)

13 Test at a 5% level of significance whether the population standard deviation of selling prices of houses in Pretoria is R55 000. Use the information for residences in Pretoria with numbers 415 to 434 in Appendix A, Data Set 1, column 3 (as in Exercise 2) and assume normality.

14 A machine that manufactures 10 cm steel bolts is set to provide a variation of less than 20.25 mm^2 in the length of these bolts. A random sample of 60 bolts gives $\bar{x} = 9.98$ cm and $s^2 = 15.17$ mm^2. Decide, on the basis of the sample information, whether it is necessary to check the setting if the lengths are assumed to be normally distributed ($\alpha = 0.01$).

15 The reaction time ($\frac{1}{100}$ second) of a random sample of 25 persons, from the point at which they see a light beginning to burn until the point at which they switch it off, was measured. The sample variance was $s^2 = 3.63$. Test whether the population variance of the reaction time for the persons was 4, assuming normality ($\alpha = 0.05$).

16 Use the information given for the first fifteen residences in Pretoria and Johannesburg respectively to test whether the mean price per square metre in the period October to December 1991 was identical for both areas

(Appendix A, Data Set 1, column 8). Let $\alpha = 0.01$. What assumptions are required?

17 Was the assumption made in Exercise 16, namely that the population variances were equal, justified? Use the F test to test at the 5% level of significance.

18 Two procedures were developed in a research laboratory to purify used oil, namely procedure A, which is quite simple and quick, and procedure B, which is considerably more expensive. Preliminary tests yielded the following results (mℓ/ℓ oil recycled):

Procedure A:	834.9	789.1	813.4	824.3	825.9
	797.7	818.0	815.7	805.4	802.5
	836.4	849.3	814.3		
Procedure B:	816.6	818.3	823.0	831.2	853.9
	841.1	834.3	824.5	841.8	829.4
	829.8	837.5	826.3	817.5	820.4

Would you say that the mean amount recycled with procedure B is more than with procedure A? Let $\alpha = 0.05$ and assume that the quantity of oil recycled is normally distributed and that the variances are equal.

19 Test whether the assumption of equal variances in Exercise 18 is justified ($\alpha = 0.01$).

20 The following data reflects the braking times of ten 1 300 cm^3 cars that are also equipped with the new braking system referred to in Exercise 3. The stopping times of ten crash stops from 100 km/h (in seconds) are the following:

3.62	3.18	3.29	3.79	3.28
3.24	3.57	3.30	3.91	3.20

Test at the 5% level of significance whether the mean braking time of the 1 300 cm^3 cars is less than that of the 1 800 cm^3 cars. What assumptions should be made before testing?

21 A manufacturer of breakfast cereal uses two makes of automatic machines to fill 500 g containers with cereal. Two independent samples provide the following information:

Machine 1	Machine 2
$n_1 = 45$	$n_2 = 55$
$\bar{x}_1 = 506.3$ g	$\bar{x}_2 = 496.7$ g
$s_1 = 12.4$ g	$s_2 = 18.5$ g

Test at a 5% level of significance whether the mean mass of the containers filled by machine 1 have the same mean mass as the containers filled by machine 2. Calculate the p-value for the test as well.

22 A botanist wants to determine whether there is a difference between the growth rate of a specific type of pine tree in the summer rainfall region and the growth rate in the winter rainfall region. In a random sample of twelve trees in the summer rainfall region the mean growth of the main trunks was 33.1 cm over a period of two years, with a standard deviation of 3.74 cm. In a random sample of 26 similar trees in the winter rainfall region the mean growth of the main trunks was 36.3 cm, with a standard deviation of 5.54 cm. Test whether the variances are equal, with $\alpha = 0.05$.

23 What conclusion can the botanist draw from the data in Exercise 22 on the assumption that the trunk growth in the summer and winter rainfall regions is normally distributed in both cases? Use the result in Exercise 22 to test at the 5% level of significance.

24 Suppose we need to compare two methods for auditing accounts by using ten pairs of auditors, paired according to years of experience and gender. The following data reflects the number of accounts audited per hour:

Pair	1	2	3	4	5	6	7	8	9	10
Method 1	97	98	59	69	59	68	87	89	79	93
Method 2	84	87	63	64	65	63	83	84	82	84

Assume that the differences between the number of accounts audited are normally distributed for each pair. Test at the 5% level of significance whether the first method will entail a saving in time (and therefore in costs).

25 Ten randomly selected students scored the following marks in the final exam in Economics and Statistics:

Student	1	2	3	4	5	6	7	8	9	10
Economics	75	63	65	88	53	84	91	59	55	62
Statistics	68	70	58	81	58	79	89	62	45	65

Can you conclude that the mean marks scored in the final exams in these subjects are identical? What assumption should be made ($\alpha = 0.01$)?

26 In Chapter 3, Exercise 27 a municipality followed a new procedure for approving building plans. The department's target was a mean of five days and a standard deviation of two days. Did the department succeed in reaching its target? Use $\alpha = 0.05$ and test

(a) whether the mean is equal to five days; and

(b) whether the standard deviation is equal to two days.

27 To determine whether a certain type of fertilizer is effective, 200 out of 500 fruit trees were not fertilized for control purposes. Of these 200, 105 showed satisfactory growth, while 225 of the 300 fertilized trees showed satisfactory growth. Can you use the sample information as a basis for concluding that the fertilizer was in fact effective? Let $\alpha = 0.05$ and calculate the p-value for this test.

28 Test whether the proportion of new residences in Pretoria is the same as in the Cape. Regard the 434 residences in Pretoria and the 406 residences in the Cape (given in Appendix A, Data Set 1, column 7) as two independent samples from the populations concerned. Let $\alpha = 0.01$ and calculate the p-values for this test.

29 Two groups of patients are selected at random in an experiment to determine the effectiveness of a new type of medication (in tablet form) for treating rheumatism. Both groups undergo identical treatment, with the following exception: The 75 patients in group 1 are given a placebo that looks exactly like the new tablet, while 86 patients in group 2 are given the new tablet. After conclusion of the treatment, 29 patients from group 1 report that they have benefited from the treatment. The corresponding figure in group 2 is 51. Are the observed results adequate proof that the medication is effective? Let $\alpha = 0.01$ and calculate the p-value for this test.

30 Use the sample information in Exercises 3 and 20 to test at the 5% level of significance whether the standard deviations of the braking times of the 1 800 cm^3 and 1 300 cm^3 cars are equal.

31 Use the information given in Exercise 21 to test for equal population variations. Take $\alpha = 0.01$ and assume normality for both populations.

32 The effectiveness of a painkiller is tested by administering it to a random sample of sixty patients, of which 48 declare after a certain period that they have experienced relief. To control these results an inactive placebo is administered to a random sample of fifty patients, and of this group, 23 declare, after the lapse of the prescribed period, that they have experienced relief. Can you say that the painkiller is really effective, that is, that more people experience relief from this than from the placebo? Test at the 5% level of significance and calculate the p-value for this test.

33 A researcher wants to determine whether there is a difference between the mean nicotine content of two types of cigarette. For a random sample of fifteen cigarettes from the first type, he finds a mean nicotine content of

27.46 milligrams, with a standard deviation of 2.64. For a random sample of 36 cigarettes from the second type he finds a mean nicotine content of 21.38 milligrams, with a standard deviation of 4.89. What can you conclude from this information if we assume that the nicotine content is normally distributed with equal variance for the two types of cigarette (α = 0.05)?

34 Is the assumption, in Exercise 33, that the nicotine content has the same variance for both types of cigarette justified on the basis of the data? Test at the 5% level of significance and repeat Exercise 33 if necessary.

35 The manufacturers of a certain type of turntable needle claim that, on average, the needle gives more than 5 000 trouble-free hours of record-playing. A consumer organization tests 45 needles and finds a mean and a standard deviation of $\bar{x}$ = 4 931 hours and s = 539 hours respectively. Would you say that this supports the manufacturers' claim? Use α = 0.05 and calculate the p-value for this test.

36 A marketer of luxury kitchen appliances (electric mixers, coffee percolators, etcetera) decides that it would be profitable to establish a branch in a large country town if more than half the housewives in the town showed interest in acquiring such appliances in the foreseeable future. In an investigation 350 housewives are selected at random and 165 of them do show interest. Do you consider it advisable to open the branch in the town? Use α = 0.01 and calculate the p-value for this test.

37 According to a medical table the average height of fifteen-year-old boys with a mass of approximately 65 kg is 176 cm. A random sample of the heights of twelve fifteen-year-old boys weighing approximately 65 kg is as follows:

175 188 180 173 165 167 182 159 177 178 181 165

If we assume that the above sample is taken from a normal population, would you say that the medical table is accurate (α = 0.05)?

38 All the house sales transactions concluded in two cities during a particular week are checked. The selling price of each house is noted. The following information is obtained:

Town A	Town B
n_1 = 53	n_2 = 35
$\bar{x}_1$ = R132 653	$\bar{x}_2$ = R148 921
s_1 = R28 361	s_2 = R33 214

Is it true that, on average, the houses in town B are cheaper than the ones in town A? Let $\alpha = 0.01$ and calculate the p-value for this test.

39 Refer to Exercise 38 and test for equality of population variances. Assume that the observations are normally distributed.

40 With reference to Exercise 38, assume that 44 of the residences sold in town A and 24 of those sold in town B have double garages. Do the two cities have an equal ratio of double garages? Let $\alpha = 0.05$ and calculate the p-value for this test.

41 An economist claims that due to higher productivity, 25% of all companies employed fewer workers in 1993 than in 1990. The Department of Manpower suspected that this proportion was as high as 25%. To investigate this claim, they selected 150 companies at random and found that 47 employed fewer workers in 1993 than in 1990. What is your conclusion? Use a 5% level of significance and calculate the p-value for this test.

42 A research report reveals that, from a sample of 346 housewives in Pretoria, 255 prefer wholewheat bread to white bread. The corresponding number for a sample of 280 housewives in Durban is 212. Does this indicate that the proportion of Pretoria housewives who buy wholewheat bread is higher than in Durban? Let $\alpha = 0.01$ and calculate the p-value for this test.

43 A milk distributor receives complaints that its litre sachets of milk contain less than a litre. To investigate these complaints, the distributor measures the quantity of milk in 35 sachets and finds a mean quantity of 0.946 litres. It is known that the machine that pours the milk has a standard deviation of 0.05 litres. Do you agree that the complaints are well founded? Use a 5% level of significance and calculate the p-value for this test.

44 The chairman of a building society claimed in a speech that the mean floor area of houses in Pretoria was 250 m^2. To investigate the claim, a random sample of forty Pretoria houses was drawn and a mean floor area of 218 m^2 and a standard deviation of 32.5 m^2 were found. Test at a 5% level of significance whether these results support the claim and calculate the p-value for this test. You may assume that the distribution of floor areas can be well approximated by a normal curve.

45 Random samples of fifty strips of carpet type A, and fifty strips of carpet type B are tested for wear and tear in the passages of a large building. The strips of type A show a mean life of 87 months and a standard deviation of

5.5 months, while the strips of type B have a mean life of 75 months and a standard deviation of 8.9 months. Use a 1% level of significance to test whether, on average, type B has a shorter life than type A and calculate the p-value for this test.

46 Twelve rounds of ammunition of type A give a mean muzzle velocity of 587 metres per second, and a standard deviation of 33 metres per second. Sixteen rounds of type B give corresponding values of 820 and 48 metres per second. If you assume that the distribution of muzzle velocities can be approximated by a normal curve and has equal variances, test at a 1% level of significance whether type A has a lower muzzle velocity than type B.

47 Refer to Exercise 46 and test if the population variances are equal at the 1% level of significance and repeat Exercise 46 if necessary.

48 The mean upper-arm length of adult males is allegedly 34 cm. On the assumption that the upper-arm length of adult males is approximately normally distributed, test the allegation on the basis of the following data obtained from a sample of size 32: $\bar{x} = 3.2$ cm; $s = 2.61$. Let $\alpha = 0.05$ and calculate the p-value for this test.

49 A car rental company evaluates two types of car in terms of fuel consumption (km/ℓ) in normal urban traffic. The following results are obtained:

	Type 1	Type 2
Sample size	$n_1 = 40$	$n_2 = 50$
Arithmetic mean	$\bar{x}_1 = 10.94$	$\bar{x}_2 = 11.43$
Standard deviation	$s_1 = 2.12$	$s_2 = 4.37$

Test if the variances are equal. Let $\alpha = 0.05$.

50 Is the fuel consumption of the second make significantly lower with respect to the data in Exercise 49? Use a level of significance $\alpha = 0.01$.

51 The manager of a marketing organization evaluates two specific approaches that can be followed when selling assurance policies. Sixty salesmen are selected at random and are randomly allocated to one of the two sales approaches. The following data represents the sales figures (number of policies) after a month:

	Approach 1	Approach 2
Sample size	$n_1 = 30$	$n_2 = 30$
Arithmetic mean	$\bar{x}_1 = 13.55$	$\bar{x}_2 = 17.26$
Standard deviation	$s_1 = 4.13$	$s_2 = 7.51$

Test whether the population standard deviations of the number of policies sold are equal ($\alpha = 0.05$).

52 Refer to the data in Exercise 51. Is there a significant difference between the mean number of policies sold for the two approaches ($\alpha = 0.05$)?

53 The ages of respondents in a marketing research investigation were divided as in the accompanying frequency table. According to a previous survey among 16 000 consumers of the product the mean age of the consumer population is 35, with a standard deviation of 12 years.

Age (years)	Class mark	Frequency
[18, 25)	21.5	14
[25, 30)	27.5	18
[30, 35)	32.5	16
[35, 45)	40.0	34
[45, 55)	50.0	27
[55, 65)	60.0	7
[65, 85)	75.5	4

(a) Calculate the mean and standard deviation in age for this sample.

(b) Is the mean age of the sample equal to that of the population? Use $\alpha = 0.01$ and calculate the p-value for this test.

(c) Is the sample standard deviation in age equal to that of the population? Use $\alpha = 0.05$.

54 Two machines at a factory that fill tins of soup are evaluated with regard to the mass of the content. The prescribed mean mass per tin is 532 g. Additional mass means a loss to the factory. The machines fill 6 000 tins per hour and the tins are tested every ten minutes. The following mass was measured during an eight-hour time shift on each of the two machines:

Machine A		Machine B	
533	534	530	535
533	533	534	534
533	534	531	534
535	532	529	531
534	534	532	532
534	534	532	532
532	533	531	535
529	530	533	531
535	534	530	531
531	537	530	533
536	534	528	535
523	533	529	531
535	535	534	533
530	538	537	532
538	534	528	532
536	534	530	529
536	531	531	532
533	534	530	528
511	532	535	532
533	530	530	533
538	531	528	530
534	530	528	529
538	536	530	534
540	533	531	532
537	535	532	534
534	538	534	530

Data obtained from MBL Group KIM00192

(a) Test whether the mean mass for machine A is equal to the specification of 532 g, with $\alpha = 0.05$.

(b) Test whether the mean mass for machine B is equal to the specification of 532 g, with $\alpha = 0.05$.

(c) Test whether the mean mass of soup tins filled by machine A is equal to the mean mass of soup tins filled by machine B ($\alpha = 0.05$).

(d) Test whether the standard deviations of the mass of the soup tins differ significantly for machines A and B.

55 In a coal mine it is important to determine whether the heat value of mined coal remains constant enough to sell it at a guaranteed minimum heat value. The higher the heat value, the better the quality of the coal and the less coal will remain after burning. The heat value is measured with the SI unit 'calorific value', CV, and is given in megajoule per kilogram. The amount of coal that is mined daily (tons) and the calorific value for one month are given in the accompanying table:

Coal produced on a daily basis in February 1993

Date	Daily quantity (ton)	Daily calorific value (CV)
1	8 862	23.57
2	4 165	21.84
3	4 287	22.31
4	4 777	22.84
5	4 383	23.03
6	1 435	23.94
7	1 858	22.82
8	8 189	24.11
9	2 673	21.46
10	4 263	22.89
11	4 604	22.49
12	2 478	23.96
13	1 942	23.87
14	1 700	23.85
15	3 289	22.89
16	8 023	23.30
17	1 502	24.61
18	7 726	23.19
19	4 425	23.30
20	3 652	22.45
21	3 183	22.80
22	5 785	23.53
23	3 679	24.30
24	4 811	21.81
25	6 115	23.39
26	5 675	22.80
27	856	23.16
28	2 385	20.94

Data obtained from MBL Group MBG00193

According to specifications the mean amount of coal mined should be 4 100 tons, and the mean calorific value 22.80. Determine whether these requirements were met during the month of February ($\alpha = 0.05$) for:

(a) the daily amount of coal mined; and

(b) the mean calorific value.

56 The effect of food supplements to optimize the condition of cattle is investigated in a research programme on an experimental farm. Mass is one of the most important measures of the condition of cattle. It is particularly important after calving, because a heavier cow is in better condition. More reserves lead to increased milk production over a longer period. In the section of the research programme set aside to investigate this aspect, specific attention was devoted to the effect of food supplements on the

condition of gravid cows in dry years. The sample comprised twenty Drakensberger and twenty Bonsmara cattle, both of which are classified as medium-frame breeds, that is, cattle with a medium build. The two groups of twenty experimental animals were paired so that the mean initial mass after conception was more or less equal. Both groups of cows were allowed to graze normally. The experimental group was given a daily supplement one month before calving, and the mass of all the cows was determined immediately after calving. The pasture had been over-grazed beforehand to simulate a dry year.

Observation pair	Control group (no mash supplement)	Experimental group (mash supplement)	Cattle breed
1	625	540	Drakensberger
2	530	530	Drakensberger
3	515	555	Drakensberger
4	530	565	Drakensberger
5	445	610	Drakensberger
6	540	545	Bonsmara
7	580	520	Bonsmara
8	395	485	Bonsmara
9	435	445	Bonsmara
10	590	615	Drakensberger
11	565	570	Drakensberger
12	550	620	Drakensberger
13	515	525	Bonsmara
14	420	515	Bonsmara
15	530	410	Bonsmara
16	455	520	Bonsmara
17	660	550	Drakensberger
18	495	540	Bonsmara
19	530	555	Drakensberger
20	510	460	Bonsmara

Data obtained from MBL Group PTA00593

According to the literature, the optimum mass of a medium-frame cow is 550 kg.

(a) Test (at $\alpha = 0.05$) whether the mean mass of the cows differs significantly from the optimum mass of 550 kg for the:

(i) Drakensberger breed in the experimental group;
(ii) Drakensberger breed in the control group;
(iii) Bonsmara breed in the experimental group; and
(iv) Bonsmara breed in the control group.

(b) Consider only the experimental group of cattle and test with $\alpha = 0.05$:

(i) whether the variance in body mass for the two cattle breeds is equal;

(ii) whether the mean body mass for the two cattle breeds is equal.

(c) Now consider the control group of cattle and test with $\alpha = 0.05$:
   (i) whether the variance in body mass for the two cattle breeds is equal;
   (ii) whether the mean body mass for the two cattle breeds is equal.

(d) Test whether the mass of the experimental group of cattle is significantly greater than that of the control group. Use $\alpha = 0.05$ and test for:
   (i) the Drakensberger breed;
   (ii) the Bonsmara breed; and
   (iii) both breeds in combination.

57 An engineer needs to determine what aggregate (dolomite or quartzite) will give the strongest concrete mixture, with due consideration to the fact that the minimum pressure resistance required is 45 MPa, and that there is no significant price difference between the two aggregates. Concrete samples with the two different aggregates were mixed over a period of twenty days and subjected to pressure tests. The other ingredients of the concrete (sand, water and cement) were kept constant to ensure that the tests would be meaningful for determining the pressure resistance of the respective aggregates.

Pressure resistance of concrete mixtures with dolomite and quartzite as aggregates

Observation	dolomite (MPa)	quartzite (MPa)
1	47.3	43.5
2	46.2	46.8
3	46.8	45.6
4	45.9	47.4
5	46.5	44.5
6	48.1	42.4
7	47.6	48.1
8	48.3	46.8
9	45.8	43.2
10	46.9	45.7
11	45.2	47.8
12	46.1	43.2
13	48.5	42.1
14	47.3	45.6
15	47.5	48.9
16	47.4	45.7
17	48.2	46.8
18	46.1	43.7
19	46.7	44.2
20	45.6	45.9

Data obtained from MBL Group JHB00293

(a) Test whether the mean pressure resistance of the dolomite mixture differs significantly from 45 MPa ($\alpha = 0.05$).

(b) Test whether the mean pressure resistance of the quartzite mixture differs significantly from 45 MPa ($\alpha = 0.05$).

(c) Test whether the dolomite mixture is significantly stronger than the quartzite mixture ($\alpha = 0.01$).

58 The ash content of the coal at a coal power station in Zimbabwe has to remain within certain specification limits. Before coal is delivered to the power station, its ash content has to be determined. The original method of determining the ash content is excessively time consuming and the information is often not ready in time. A new, quicker method for determining the ash content has been developed. However, the accuracy of the new method is still uncertain. A number of samples of coal were taken and the ash content was determined by means of the old method and the new, quick method.

Observation	Ash content	
	Old method	New method
1	20.5	20.9
2	22.0	22.2
3	22.4	22.5
4	22.5	22.8
5	22.6	23.6
6	23.2	23.7
7	23.3	24.1
8	23.4	24.1
9	23.8	24.2
10	23.9	24.2
11	24.3	24.4
12	24.3	24.6
13	24.3	24.6
14	24.7	24.8
15	24.7	25.1
16	25.0	25.2
17	25.0	25.4
18	25.2	25.4
19	25.2	26.0
20	25.2	26.2
21	25.6	26.6
22	26.4	27.0
23	26.9	27.0
24	27.2	28.0
25	27.3	29.7
26	29.4	29.9

Use $\alpha = 0.05$ to test whether the quick method yields a higher measurement than the old method.

59 Use Appendix A, Data Set 5, to test whether the mean electricity consumption of the brewery for 1992 differs from the one for 1993. Use $\alpha = 0.01$ and calculate the p-value for this test.

60 Use Appendix A, Data Set 5, to test whether the mean amount of beer produced in the brewery in 1992 differs from the amount for 1993. Use $\alpha = 0.01$ and calculate the p-value for this test.

61 Use Appendix A, Data Set 5, to test whether the mean amount of electricity consumed in 1992 to produce 1 hectolitre of beer differs from the amount for 1993. Use $\alpha = 0.01$ and calculate the p-value for this test.

62 A factory produces two thicknesses of paper-isolated lead cables. The specification limits for the thickness of the outer layer of lead for the two types of cables are:

35 mm cable: 1.54 ± 0.04
185 mm cable: 2.05 ± 0.05

During the manufacturing process measurements were taken from both types of cable to determine whether the above specifications were met. During manufacturing, the cables are cut to specific lengths before the lead is added. The radial thickness of the cables is measured once after the cable has undergone the process during which lead is added. The end of the cable is known as the inside end (that is, the end that will be on the inside of a drum on which the cable will be wound). The other end of the cable, which is the last to go through the production process, is known as the outside end. The radial thickness is measured again at the outside end. Because of the weight of the cable the first section runs through the process at a slower rate than the last section during the manufacturing process. The data for twenty samples of the 35 mm cable and sixteen samples of the 185 mm cable is given in the accompanying table.

(a) The production manager suspects that the mean amount of lead added to the cables is significantly more than the top specification limit, which means that the company could save on production costs. Use $\alpha = 0.05$ and test whether his suspicion is valid on:

(i) the inside end of the 35 mm cable;
(ii) the outside end of the 35 mm cable;
(iii) the inside end of the 185 mm cable; and
(iv) the outside end of the 185 mm cable.

(b) Furthermore, the production manager suspects that the mean lead thickness on the inside end of the lead cable is significantly more than the mean thickness of the lead on the outside end. Use $\alpha = 0.05$ and test whether his suspicion is valid for:

(i) the 35 mm cable; and
(ii) the 185 mm cable.

35 mm cable			185 mm cable		
Sample	Inside end	Outside end	Sample	Inside end	Outside end
1	1.91	1.86	1	2.28	2.17
2	1.75	1.74	2	2.27	2.14
3	1.91	1.77	3	2.30	2.14
4	1.91	2.10	4	2.56	2.59
5	1.83	1.86	5	2.69	2.44
6	1.86	1.80	6	2.57	2.42
7	1.67	1.87	7	2.58	2.41
8	1.97	1.92	8	2.34	2.10
9	1.96	1.96	9	2.27	2.30
10	1.87	1.70	10	2.44	2.38
11	1.85	1.74	11	2.45	2.41
12	1.85	1.71	12	2.29	2.31
13	1.75	1.77	13	2.27	2.42
14	1.82	1.73	14	2.29	2.12
15	1.84	1.77	15	2.21	2.10
16	1.94	1.75	16	2.38	2.16
17	2.24	2.06			
18	2.08	2.05			
19	2.02	1.89			
20	2.03	1.91			

Data obtained from MBL Group PEL00292

63 A restaurant studies the dining-out habits of married couples. Of a sample of 300 couples, 42% of the wives held full-time jobs. In this group 56 indicated that they dined out at least once a week. In the group in which the wives did not work full-time, 65 indicated that they dined out at least once a week. Can the restaurant conclude that couples of which the wives worked full-time dined out more regularly than couples of which the wives did not work full-time ($\alpha = 0.05$)? Calculate the p-value for this test as well.

64 An advertising company is conducting a survey among television viewers to determine whether they like a certain advertisement. A random sample yields the following data:

Gender	Opinion		
	Favourable	Neutral	Unfavourable
Male	189	95	190
Female	109	93	149

Test the hypothesis that the proportion of men who enjoyed the advertisement is greater than the proportion of women. Use $\alpha = 0.01$ and calculate the p-value for this test.

65 The following hypotheses were tested in Exercise 4, with $\alpha = 0.05$:

H_0: $\mu = 85$ against H_1: $\mu > 85$.

Calculate and draw the power function $P_R(\mu_1)$ for $\mu_1 = 86, 87, 88, 90, 92, 95$ and 100 if we assume that $\sigma = 5.68$.

66 In Exercise 6 the mean vocabulary of children of a specific age is given as 1 040 words. An educational researcher has tested the following hypotheses:

H_0: $\mu = 1\,040$ against H_1: $\mu > 1\,040$.

After implementation of a special teaching programme to expand the vocabulary of children in this age group, 120 children were tested and it was found that $\bar{x} = 1\,086$. Assume that $\sigma = 145.3$, and calculate and draw the power function $P_R(\mu_1)$ for $\mu_1 = 1\,045, 1\,050, 1\,055, 1\,065, 1\,080, 1\,100, 1\,150$ and 1 200.

67 In Exercise 7 a penicillin manufacturer claimed that less than 12% of all people are allergic to its product. The relevant hypotheses were H_0: $\pi = 0.12$ against H_1: $\pi < 0.12$, with $\alpha = 0.01$. To test the claim, a random sample of 90 persons was selected and they were treated with penicillin. Calculate and draw the power function $P_L(\pi_1)$ for $\pi_1 = 0, 0.03, 0.05, 0.07, 0.08, 0.09, 0.10$ and 0.11.

68 In Exercise 8 a soft drink company considered changing from 500 ml glass containers to 500 ml cardboard containers. The relevant hypotheses were H_0: $\pi = 0.75$ against H_1: $\pi > 0.75$, with $\alpha = 0.01$. An opinion poll among 260 consumers showed that 196 of them were in favour of the proposed change. Calculate and draw the power function $P_R(\pi_1)$ for $\pi_1 = 0.75, 0.76, 0.77, 0.78, 0.80, 0.85, 0.90$ and 1.0.

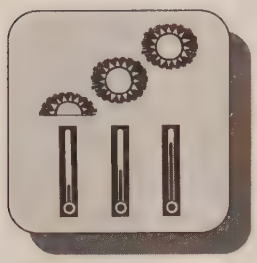
69 In Exercise 9 a manufacturer of car parts claimed that less than 10% of all parts supplied to an assembly plant did not comply with the specifications. When 235 parts were tested 19 were found to be defective. For a 5% level of significance, calculate and draw the power function $P_L(\pi_1)$ for $\pi_1 = 0, 0.03, 0.05, 0.07, 0.08, 0.090, 0.095$ and 0.10.

70 In Exercise 11, 135 persons from a random sample of 250 persons in Cape Town indicated that they watched the news on television regularly. The relevant hypotheses were H_0: $\pi = 0.70$ against H_1: $\pi \neq 0.70$, with

$\alpha = 0.05$. Calculate and draw the power function $P_T(\pi_1)$ for $\pi_1 = 0, 0.50, 0.55, 0.60, 0.65, 0.67, 0.69, 0.70, 0.71, 0.73, 0.75, 0.80, 0.85$ and 0.90.

71 In Exercise 14 a machine that manufactures 10 cm steel bolts was set so that the variance in the length of these bolts would be less than 20.25 mm^2. If $\alpha = 0.01$, calculate and draw the power function $P_L(\sigma_1^2)$ for $\sigma_1^2 = 16.00, 17.00, 18.00, 19.00, 19.05, 19.75, 20.0$ and 20.25 mm^2.

72 In Exercise 43 a milk distributor received complaints that the litre sachets of milk contained less than a litre. To investigate these complaints, the distributor measured the milk in 35 sachets and found a mean content of 0.946 litres. It is known that the machine that pours the milk has a standard deviation of 0.05 litres. If a 5% level of significance is used, calculate and draw the power function $P_L(\mu_1)$ for $\mu_1 = 0.80, 0.85, 0.90, 0.925, 0.95, 0.975, 0.99$ and 1.0.



CHAPTER 13

Linear regression and correlation

1 Introduction

In Chapter 6 we looked at situations where two or more characteristics of a particular population are simultaneously considered. In order to obtain more information on **two** characteristics, a sample of n observation pairs (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) is taken from the population under consideration. As indicated in Chapter 6, there is often some or other relationship between x and y that can be adequately described by a mathematical equation $\hat{y} = f(x)$. A scatter diagram of the relevant data suggests the nature of this relation. In this chapter we discuss, for the bivariate case, only the statistical inference based on the assumption that the association between x and y can be adequately described by a straight line. As in Chapter 6, we assume throughout that y is the **dependent** variable which is explained by the **independent** variable x.

Suppose x and y are valuations and selling prices of residences in Johannesburg during the period October to December 1991. Table 6.1 shows the ten pairs of observations regarding residences in Johannesburg as contained in Appendix A, Data Set 1, observations 521 to 530. The scatter diagram in Figure 6.1 shows an almost linear relationship between x and y. In Example 6.3 a straight line with the equation

$$\hat{y} = -1.1360 + 1.0599x \quad (x\text{- and } y\text{-units R1 000})$$

is fitted to the data and in Example 6.5 we see that the correlation coefficient $r = 0.99$. Since other samples of size 10 from the same population may yield data sets that differ substantially from those in Table 6.1 and also from one another, it is logical that the values of the intercept a, the slope b and the correlation coefficient r will also vary from one sample to the next. Example 13.1 shows this very clearly.

EXAMPLE 13.1

In order to further investigate the relationship between valuation, x, and selling price, y, of houses in Johannesburg, five additional samples (each of size 10) are taken from Appendix A, Data Set 1. These samples, together with the respective averages of x and y are shown in Table 13.1.

Table 13.1 Five samples (x_i, y_i), $i = 1, 2, \ldots, 10$ from the population of Johannesburg houses, where x represents house valuations and y selling prices (R1 000)

Sample 1: Houses 531–540

x	74.0	101.0	79.0	218.5	119.0	88.0	47.5	142.0	134.0	85.0	$\bar{x} = 108.80$
y	80.5	110.0	84.0	250.0	130.0	93.0	50.0	165.0	178.0	93.0	$\bar{y} = 123.35$

Sample 2: Houses 541–550

x	85.0	313.0	102.0	100.0	98.0	70.5	119.0	91.0	129.0	395.0	$\bar{x} = 150.25$
y	90.8	415.0	150.0	107.0	118.0	100.0	70.0	100.0	140.0	430.0	$\bar{y} = 172.08$

Sample 3: Houses 551–560

x	320.0	138.0	135.0	219.0	72.0	186.0	75.0	462.0	84.0	102.0	$\bar{x} = 179.30$
y	370.0	145.0	140.0	240.0	80.0	200.0	85.0	462.0	104.0	106.0	$\bar{y} = 193.20$

Sample 4: Houses 561–570

x	183.0	121.0	69.0	83.5	119.3	221.0	174.0	244.0	346.0	249.0	$\bar{x} = 180.98$
y	220.0	140.0	75.0	89.4	126.3	225.0	181.7	265.0	405.0	285.0	$\bar{y} = 201.24$

Sample 5: Houses 571–580

x	189.0	345.0	127.6	140.0	158.0	208.0	281.0	141.4	229.5	211.5	$\bar{x} = 203.10$
y	191.0	390.0	130.6	165.0	165.0	213.0	320.0	144.4	233.5	216.0	$\bar{y} = 216.85$

To determine the least squares values of the intercept a and the slope b and also the correlation coefficient r (if a linear relationship between x and y is assumed), we use the following formulas already derived in Chapter 6:

$$b = \frac{\sum xy - \frac{1}{n}\sum x \sum y}{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}$$

$$r = \frac{b\sqrt{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}}{\sqrt{\sum y^2 - \frac{1}{n}\left(\sum y\right)^2}}$$

$$a = \bar{y} - b\bar{x}$$

Table 13.2 shows the different values of a, b and r as calculated for each of the bivariate data sets given in Table 13.1. From this table it is clear that the values of a, b, and r vary from one sample to the next. We can therefore consider a, b and r, as calculated for an observed data set, as observed values of the stochastic variables A, B and R respectively. Note, however, that R here has a different meaning than in Chapter 6 where R^2 denoted the coefficient of determination.

Table 13.2 Calculated values of a, b, and r with regard to the five samples (x_i, y_i), $i = 1, 2, \ldots, 10$ given in Table 13.1

Sample number	Intercept a	Slope b	Correlation coefficient r
1	–9.2439	1.2187	0.9887
2	–4.9020	1.1779	0.9697
3	9.7020	1.0234	0.9941
4	–9.1362	1.1624	0.9930
5	–20.7608	1.1699	0.9901

The stochastic variables A, B and R are respectively estimators of the population intercept α, the population slope β and the population correlation coefficient ρ between the two relevant population characteristics. In order to simplify the statistical inference regarding the intercept α and slope β the following assumption is made:

> The values of the independent or explanatory variable x is fixed and can be predetermined. The dependent variable Y is a stochastic variable so that, for any given value of x, it follows that
>
> $$\mu_y = \alpha + \beta x.$$

Statistical inference regarding α and β is based on a single observed sample $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ from which the observed **sample regression line of y given x**, namely $\hat{y} = a + bx$ is calculated. The line $\hat{y} = a + bx$ is an estimate of the **population regression line of y given x**, namely $\mu_y = \alpha + \beta x$.

> The statistical inference concerned with α, β and $\alpha + \beta x$ is known as **bivariate linear regression theory**.

Notes

(a) It can easily be shown in terms of the above assumption that A and B are unbiased estimators of the population parameters α and β while $A + Bx$ is an unbiased estimator of the population parameter $\alpha + \beta x$.

(b) The assumption that the values of the independent variable x can be regarded as fixed or given values, while the dependent variable Y is a stochastic variable, is essential in regression theory. However, this

assumption can usually be justified in practical situations. When deciding on the price he/she is prepared to pay, a prospective buyer will take into consideration, among other things, the given house valuation of a specific house which is for sale. However, since there are many factors involved in determining the price of a house not every buyer of a house that has been valued for say R220 000, will necessarily pay the same amount. Although the house valuation x can therefore be regarded as fixed or given, the selling price Y is a stochastic variable. In the case of time series data (see Chapter 7) the time series x are fixed preselected points in time (for example months or years) while the Y values (for example: production or rainfall) are clearly stochastic variables.

2 Sample variance around the regression line

Figure 13.1 shows the scatter diagram of the second sample of Example 13.1. From Table 13.2 it follows that, for this bivariate data set, the equation of the least squares regression line of y given x is given by

$$\hat{y} = -4.9020 + 1.1779x.$$

This line is also shown in Figure 13.1 and it appears to be a good fit. This is further confirmed by the fact that $r = 0.9697$. The degree of deviation of the observed y values from the sample regression line is reflected in the following measure:

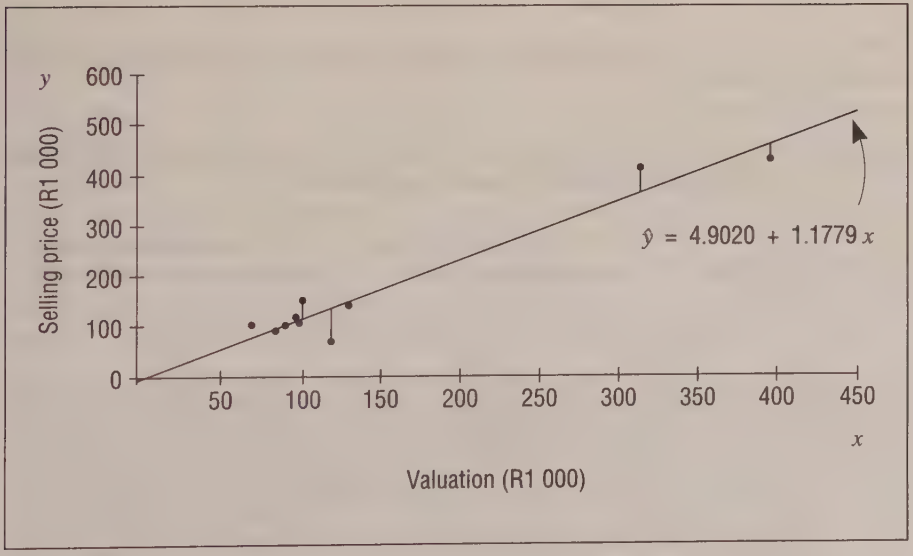


Figure 13.1 Representation of the variance around the regression line $\hat{y}$ for the second data set in Table 13.1

The **sample variance around the regression line** $\hat{y}$ is a measure of spread of the observed y values around the regression line and is defined by

$$s^2_{y\cdot x} = \frac{\sum (y - \hat{y})^2}{n - 2}$$

If $\hat{y} = a + bx$ is substituted in the definition of $s^2_{y\cdot x}$ it follows after simplification that

$$s^2_{y\cdot x} = \frac{\sum y^2 - a\sum y - b\sum xy}{n - 2}$$

It is self-evident that $s^2_{y\cdot x}$ will vary from one sample to the next so that $s^2_{y\cdot x}$ can be regarded as the observed value of the stochastic variable $S^2_{Y\cdot x}$.

EXAMPLE 13.2

Consider once again the data given in Example 13.1. For the first sample we obtain

$\Sigma y = 1\,233.50$; $\Sigma y^2 = 183\,743.25$; $\Sigma xy = 159\,544.00$; $a = -9.243868$; $b = 1.218694$ and $n = 10$

so that it follows that

$$\begin{aligned} s^2_{y\cdot x} &= \tfrac{1}{8}\{\Sigma y^2 - a\Sigma y - b\Sigma xy\} \\ &= \tfrac{1}{8}\{183\,743.25 - (-9.243868 \times 1\,233.50) - 1.218694 \times 159\,544.00\} \\ &\approx 88.78 \end{aligned}$$

Table 13.3 gives a summary of the $s^2_{y\cdot x}$ values (rounded to 2 decimals) for all five samples in Example 13.1.

Table 13.3 Values of the sample variance around the regression line with regard to the five samples in Example 13.1.

Sample number	Sample variance around the regression line $s^2_{y\cdot x}$	Sample standard deviation around the regression line $s_{y\cdot x}$
1	88.78	9.42
2	1 207.04	34.74
3	222.60	14.92
4	159.08	12.61
5	147.41	12.14

If a straight line $\hat{y} = a + bx$ is fitted to a bivariate data set by means of the least squares method, the coefficient of determination simplifies to r^2 (see Chapter 6, Section 4), that is

$$r^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

Now substitute the sample variance of the y values, s_y^2, and the variance around the regression line, $s_{y \cdot x}^2$, in the above equation. It then follows that

$$r^2 = 1 - \frac{(n - 2)s_{y \cdot x}^2}{(n - 1)s_y^2}$$

$$\approx 1 - \frac{s_{y \cdot x}^2}{s_y^2} \text{ for } n \text{ large}$$

A small observed value of $s_{y \cdot x}^2$ relative to s_y^2 indicates that the regression line of y given x explains the variation in y well. Consequently, information on x would, in this case, lead to a good estimate or prediction of the corresponding y value. In such a case the value of r^2 is relatively large. On the other hand, a large observed value of $s_{y \cdot x}^2$ relative to s_y^2 will cause the value of r^2 to be relatively small. This again indicates that a linear function of x presents no substantial explanation of y.

Note

It can be shown that the sample variance $s_{y \cdot x}^2$ is an unbiased estimator of $\sigma_{y \cdot x}^2$ = the variance around the population regression line $\mu_y = \alpha + \beta x$.

3 Inference regarding the intercept α

In practice it is often necessary to compile, on the basis of the sample information, a confidence interval for α or to test H_0: $\alpha = \alpha_0$ against the appropriate alternative. In Section 1 we pointed out that the stochastic variables A and B are unbiased estimators of α and β. In this and in the next section we define the pivotal quantities associated with the estimators A and B. As in Chapter 11 and 12, these quantities are used to compile confidence intervals and to solve hypothesis testing problems regarding α and β. The statistical inference regarding α and β is based on a very important assumption, namely

Inference assumption regarding linear regression: We assume that $Y_1, Y_2, \ldots, Y_n$ are independent stochastic variables that correspond respectively with the given values $x_1, x_2, \ldots, x_n$. We also assume that Y_i has a $n(\alpha + \beta x_i, \sigma^2)$ distribution.

These important assumptions will often be referred to in the following sections. With regard to the intercept α the following applies:

If the inference assumption for linear regression is valid, the **pivotal quantity** regarding α is defined by

$$T_A = \frac{A - \alpha}{S_A}$$

$$\text{with } S_A = S_{y \cdot x} \sqrt{\frac{\sum x^2}{n\left\{\sum x^2 - \frac{1}{n}(\sum x)^2\right\}}}$$

in which T_A has a t$(n-2)$ distribution.

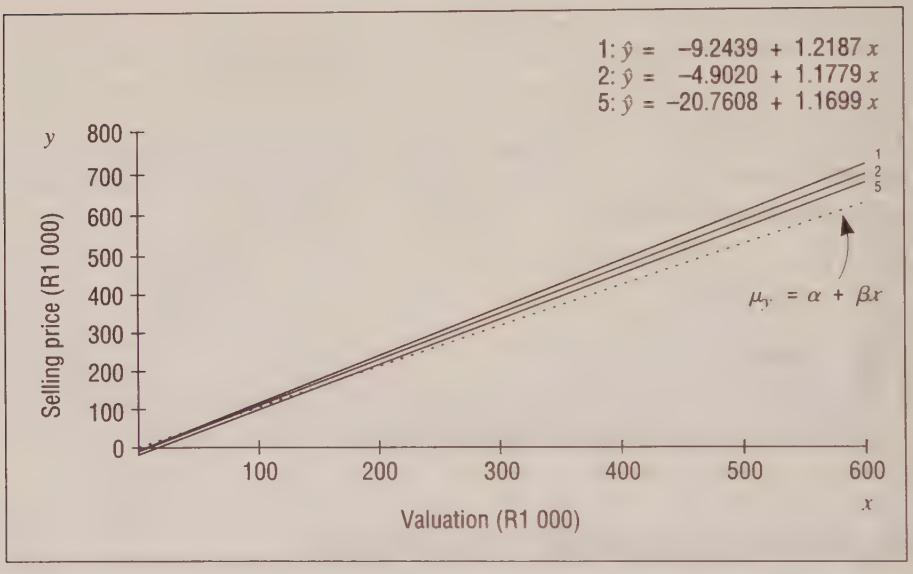


Figure 13.2 Graphic representation of the population regression line $\mu_y = \alpha + \beta x$ and of a number of observed sample regression lines, each of the form $\hat{y} = a + bx$

Figure 13.2 is a graphic representation of the sample regression lines of y given x for samples 1, 2 and 5 of Example 13.1. The 'population regression line' $\mu_y = \alpha + \beta x$ is also indicated in this figure. The values of α and β are obviously unknown but have been determined here for purposes of illustration as the a and b values of the population of residences sold in Johannesburg as shown in Appendix A, Data Set 1. This figure shows that the estimates of the intercept vary considerably around the unknown population intercept α. To compile a confidence interval for α, we use the fact that T_A has a t$(n-2)$ distribution. Consequently it follows that

$$P\left(-t_{n-2,\,0.025} < \frac{A - \alpha}{S_A} < t_{n-2,\,0.025}\right) = 0.95$$

Using this equation, a 95% confidence interval for α can be determined in a method analogous to the one discussed in Chapter 11.

Under the inference assumption regarding linear regression and on the basis of the observed sample (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), the **95% confidence interval for** α is given by

$$(a - t_{n-2,\,0.025} \times s_a,\ a + t_{n-2,\,0.025} \times s_a)$$

with s_a the calculated value of S_A.

EXAMPLE 13.3

It is common knowledge that it pays to advertise. An investigation is undertaken to determine the effect of advertisement on sales in pharmacies. Twelve pharmacies are selected at random and both the advertising expenditure and the business profit are determined. The data is reflected in Table 13.4.

Table 13.4 Advertising expenditure and business profit for twelve pharmacies

Pharmacy	Advertising expenditure (R) x	Business profit (R) y
1	1 581.61	6 302.52
2	2 292.80	6 901.88
3	2 595.47	7 773.93
4	1 062.19	4 527.45
5	1 930.86	5 720.73
6	655.12	4 660.12
7	1 294.00	6 507.83
8	1 232.03	5 348.99
9	2 096.78	7 578.83
10	1 381.22	5 125.67
11	1 104.78	5 258.17
12	403.62	4 445.21

Suppose we wish to calculate a 95% confidence interval for α. For this sample

$\Sigma y^2 = 424\,815\,913.3$, $\Sigma y = 70\,151.33$, $\Sigma xy = 110\,463\,001.8$,
$a = 3\,561.978974$ and $b = 1.554557$

so that $s^2_{y\cdot x} = \frac{1}{10}\{424\,815\,913.3 - (3\,561.978974 \times 70\,151.33) - (1.554557 \times 110\,463\,001.8)\}$

$= 321\,738.5248$

so that $s_{y\cdot x} = 567.22$

Furthermore $\Sigma x = 17\,630.48$ and $\Sigma x^2 = 30\,660\,573.63$

$$s_a = s_{y \cdot x}\sqrt{\frac{\sum x^2}{n\left\{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2\right\}}}$$

$$= 567.22\sqrt{\frac{30\,660\,573.63}{12\left\{30\,660\,573.63 - \frac{1}{12}(17\,630.48)^2\right\}}}$$

$$= 415.67$$

Since it follows from Table B.3 that $t_{10,\,0.025} = 2.228$, the 95% confidence interval for α is given by

$$(3\,561.98 - 2.228 \times 415.67, 3\,561.98 + 2.228 \times 415.67)$$

that is $(2\,635.87, 4\,488.09)$

From this data we can deduce that, where a pharmacy has no advertising expenses, its trading profit will lie between approximately R2 600 and R4 500.

The procedure for testing H_0: $\alpha = \alpha_0$ is explained in Table 13.5. In this chapter we shall indicate the level of significance of a statistical test by ε instead of by the commonly used α, in order to prevent confusion.

Table 13.5 Testing H_0: $\alpha = \alpha_0$ on the basis of a bivariate sample $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ in which Y_i has a $n(\alpha + \beta x_i, \sigma^2)$ distribution for a given x_i

Test statistic	Distribution of test statistic under H_0	Table with critical values
$T_A = \dfrac{A - \alpha_0}{S_A}$ with $S_A = S_{y \cdot x}\sqrt{\dfrac{\sum x^2}{n\left\{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2\right\}}}$ and $S_{y \cdot x} = \sqrt{\dfrac{\sum Y^2 - A\sum Y - B\sum xY}{n-2}}$	$t(n-2)$	Table B.3
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
H_1: $\alpha \neq \alpha_0$	$t \leq -t_{n-2,\,\varepsilon/2}$ or $t \geq t_{n-2,\,\varepsilon/2}$	$2P(T \geq \lvert t \rvert)$
H_1: $\alpha > \alpha_0$	$t \geq t_{n-2,\,\varepsilon}$	$P(T \geq t)$
H_1: $\alpha < \alpha_0$	$t \leq -t_{n-2,\,\varepsilon}$	$P(T \leq t)$

EXAMPLE 13.4

A manufacturer of medical equipment has developed an affordable device to measure blood-sugar levels. This device enables diabetics to determine their insulin requirements at home. To evaluate the accuracy of the device, the blood-sugar levels of fifteen people are determined by means of the apparatus and by laboratory tests. The following readings are obtained:

Table 13.6 Blood-sugar levels (millimol per litre) for fifteen patients with calibrated home and laboratory equipment

	Home equipment y	Laboratory equipment x
1	3.2	3.6
2	3.3	4.0
3	4.3	4.1
4	5.1	4.6
5	4.8	4.7
6	4.9	4.7
7	5.2	5.3
8	5.2	5.5
9	5.7	5.5
10	6.2	6.3
11	5.2	6.3
12	6.2	6.5
13	7.3	6.8
14	7.8	7.6
15	6.9	7.1

Suppose we want to test the hypothesis H_0: $\alpha = 0$, namely that the population regression line passes through its point of origin. From the above table one can calculate that $\Sigma y^2 = 464.91$, $\Sigma y = 81.3$, $\Sigma xy = 468.94$, $\Sigma x = 82.6$, $\Sigma x^2 = 475.74$, $a = -0.181215$ and $b = 1.017170$ so that the variance around the regression line $s^2_{y \cdot x} = \frac{1}{13}\{464.91 + 0.181215 \times 81.3 - 1.017170 \times 468.94\} = 0.2039$ and $s_{y \cdot x} = 0.4516$.

The steps for testing the hypothesis are:

Step 1

H_0: $\alpha = 0$
H_1: $\alpha \neq 0$

Step 2

Let the level of significance $\varepsilon = 0.05$. Under H_0 $T_A = \dfrac{A - 0}{S_A}$ has a t(13) distribution provided the inference assumption for linear regression is valid.

Step 3

Reject H_0 if $t_a \leq -2.160$ or $t_a \geq 2.160$ or if the p-value is smaller than 0.05.

Step 4

$$s_a = s_{y \cdot x} \sqrt{\frac{\sum x^2}{n\left\{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2\right\}}}$$

$$= 0.4516 \sqrt{\frac{475.74}{15\left\{475.74 - \frac{1}{15}(82.6)^2\right\}}}$$

$$= 0.556$$

$$t_a = \frac{-0.181 - 0}{0.556}$$

$$= -0.326$$

Step 5

Since $-2.160 < t_a < 2.160$, H_0 is not rejected and the assumption that the population regression line passes through the origin appears to be realistic. The p-value = 0.7499.

4 Inference regarding the slope β

As mentioned in Chapter 6, the slope of the regression line is always of practical importance. Within the interval of observed x values the slope indicates the expected change in the dependent variable if the independent variable increases by one unit.

If the inference assumption for linear regression applies, the **pivotal quantity** regarding β is defined by

$$T_B = \frac{B - \beta}{S_B}$$

$$\text{with } S_B = \frac{S_{y \cdot x}}{\sqrt{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}}$$

where T_B has the t$(n-2)$ distribution.

From Figure 13.2 it is evident that the slopes of the sample regression lines of y given x vary considerably around the population slope β. To obtain an indication of the real value of β, a confidence interval is calculated for β.

On the basis of the observed sample $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and in terms of the inference assumption for linear regression, a **95% confidence interval for β** is given by

$$(b - t_{n-2,\,0.025}\, s_b,\; b + t_{n-2,\,0.025}\, s_b)$$

where s_b is the calculated value of S_B.

EXAMPLE 13.5

Just as snakes shed their skins in a process known as sloughing, lobsters also undergo a moulting process. The Cape rock lobster is said to shed moult equivalent to about a third of its skeletal mass each year. The following data gives the total amount of dry moult (in grams) and the skeletal mass (in grams) of eleven lobsters.

Table 13.7 Moult (gram) and skeletal mass (gram) of eleven lobsters

Lobster	Moult y	Skeletal mass x
1	39.1	140.0
2	51.9	147.9
3	28.2	173.4
4	87.6	284.2
5	89.5	308.2
6	111.4	356.0
7	99.9	384.0
8	148.6	467.6
9	123.5	483.5
10	155.3	563.8
11	169.2	607.9

Suppose we wish to calculate a 95% confidence interval for β on the basis of the data contained in Table 13.7. For this sample it was found that $\Sigma y^2 = 133\,172.58$, $\Sigma y = 1\,104.20$, $\Sigma xy = 468\,152.14$, $\Sigma x = 3\,916.5$, $\Sigma x^2 = 1\,661\,325.71$, $a = 0.31271$ and $b = 0.28106$, so that the sample variance around the regression line $s^2_{y\cdot x} = \frac{1}{9}\{133\,172.58 - 0.31271 \times 1\,04.20 - 0.28106 \times 468\,152.14\}$ $= 138.71613$ and $s_{y\cdot x} = 11.78$.

$$s_B = \frac{s_{y\cdot x}}{\sqrt{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}}$$

$$= \frac{11.78}{\sqrt{1\,661\,325.71 - \frac{1}{11}(3\,916.5)^2}}$$

$$= 0.0228$$

Since $t_{9,\,0.025} = 2.262$ it follows that a confidence interval for β is given by

$$(0.2811 - 2.262 \times 0.0228, 0.2811 + 2.262 \times 0.0228)$$

$$= (0.2295, 0.3327).$$

The statement regarding the amount of moulting material lost thus appears to be realistic.

Statements about the population slope sometimes have to be tested statistically. The procedure for testing H_0: $\beta = \beta_0$ is set out in Table 13.8. Example 13.6 illustrates the testing procedure.

Table 13.8 Testing H_0: $\beta = \beta_0$, on the basis of a bivariate sample $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ in which, for a given x_i, Y_i has a $n(\alpha + \beta x_i, \sigma^2)$ distribution

Test statistic	Distribution of test statistic under H_0	Table with critical values
$T_B = \dfrac{B - \beta_0}{S_B}$ with $S_B = \dfrac{S_{y \cdot x}}{\sqrt{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}}$ and $S_{y \cdot x} = \sqrt{\dfrac{\sum Y^2 - A\sum Y - B\sum xY}{n-2}}$	$t(n-2)$	Table B.3
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
H_1: $\beta \neq \beta_0$	$t \leq -t_{n-2,\,\varepsilon/2}$ or $t \geq t_{n-2,\,\varepsilon/2}$	$2P(T \geq \lvert t \rvert)$
H_1: $\beta > \beta_0$	$t \geq t_{n-2,\,\varepsilon}$	$P(T \geq t)$
H_1: $\beta < \beta_0$	$t \leq -t_{n-2,\,\varepsilon}$	$P(T \leq t)$

EXAMPLE 13.6

The public is often confronted with advertisements concerning gadgets that lay claim to effecting considerable savings in the fuel consumption of motor cars. According to the advertisers, one such device can easily be installed by the owner and offers a saving of more than 15% on fuel consumption. To test this claim scientifically, twelve motor cars of different makes and piston capacity are chosen. The distance (in km) travelled per litre of petrol, with and without the device, is reflected in the data given in Table 13.9.

The following results can be calculated from Table 13.9: $\Sigma y^2 = 1\,696.68$, $\Sigma y = 140.45$, $\Sigma xy = 1\,400.82$, $\Sigma x = 116.14$, $\Sigma x^2 = 1\,159.37$, $a = 0.33829$ and $b = 1.174370$ so that the sample variance around the regression line $s^2_{y \cdot x} = \frac{1}{13}\{1\,696.68 - 0.33829 \times 140.45 - 1.174370 \times 1\,400.82\} = 0.4096$ and $s_{y \cdot x} = 0.64$.

The claim that the new device offers a saving of more than 15% can be described as follows in terms of x and y: $\frac{y}{x} \geq 1.15$, in other words, $y \geq 1.15\,x$. We therefore test the hypotheses in Step 1.

Step 1

H_0: $\beta = 1.15$
H_1: $\beta > 1.15$

Table 13.9 Distance travelled by twelve cars with and without the fuel-saving device

Vehicle	With device y	Without device x
1	9.11	7.63
2	8.24	7.20
3	13.47	11.05
4	13.43	10.95
5	9.98	8.45
6	14.50	11.25
7	9.10	7.60
8	12.67	10.68
9	10.96	9.07
10	11.74	9.46
11	12.74	9.69
12	14.51	13.11

Step 2

Suppose the level of significance is $\varepsilon = 0.05$. If H_0 is true and the inference assumption for linear regression is met, then $T_B = \dfrac{B - 1.15}{S_B}$ has a t(10) distribution.

Step 3

Since $t_{10,\,0.05} = 1.813$, it follows that H_0 will be rejected if $t_b \geq 1.813$ or if the p-value ≤ 0.05.

Step 4

$$s_b = \frac{s_{y \cdot x}}{\sqrt{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}}$$

$$= \frac{0.64}{\sqrt{1159.37 - \frac{1}{12}(116.14)^2}}$$

$$= 0.108$$

$$t_b = \frac{1.1744 - 1.15}{0.108}$$

$$= 0.226$$

Step 5

Since $t_b < 1.813$, H_0 is not rejected. The claim that the device brings about a saving of more than 15% in fuel consumption is thus not substantiated by the findings in the sample. The p-value is 0.41288.

5 Confidence band for $\alpha + \beta x$

The Department of Flight Operations of the South African Airways is concerned about the number of landings averaged by a set of main landing tyres on the Boeing 737. The Department of Flight Training is asked to investigate the situation, with specific emphasis on the effect of the landing speed on the number of landings that can be expected from a set of tyres. A specific aircraft is set aside for one year for testing purposes, and the landing speed (independent variable x) and the number of landings (dependent variable y) are noted each month (see Example 12.17). Suppose Flight Operations wishes to predict the number of landings that can be expected from a set of tyres if the aircraft lands at a particular speed, for instance, 125 knots. The flight engineer will then have to estimate the value of $\mu_y = \alpha + \beta \times 125$. This is a typical example of a situation in which it is important to derive, from a given x value, the confidence interval for the population regression line $\mu_y = \alpha + \beta x$.

Under the inference assumption regarding linear regression and on the basis of the observed sample $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the **95% confidence interval for $\alpha + \beta x$**, in which x is specified, is given by

$$(a + bx - t_{n-2,\,0.025}\,d_x,\ a + bx + t_{n-2,\,0.025}\,d_x)$$

with d_x the observed standard deviation of $A + Bx$, namely

$$d_x = s_{y \cdot x}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}}$$

Although the confidence interval for $\alpha + \beta x$ is often compiled for the x values that correspond to the ones included in the sample, the above result is generally also valid for any other x value inside or outside the interval of observed x values for which the linear relationship is expected to remain valid. The graphic representation of the confidence intervals for $\alpha + \beta x$, based on a large set of divergent x values, leads to the so-called **confidence band for the population regression line** $\mu_y = \alpha + \beta x$. Figure 13.3 shows such a confidence band.

EXAMPLE 13.7

In an investigation into the effect of the landing speed on the number of landings made with a set of main landing tyres on the Boeing 737, the following data is obtained:

Table 13.10 Number of landings and landing speed on a set of tyres on the Boeing 737

Observation	Number of landings y	Landing speed x
1	84	112
2	85	114
3	84	116
4	80	118
5	81	120
6	76	122
7	78	124
8	71	126
9	72	128
10	68	130
11	69	132
12	64	134
$\bar{x} = \frac{1476}{12} = 123.0$	$\Sigma x^2 = 182\,120$	$\Sigma xy = 111\,632$
$\bar{y} = \frac{912}{12} = 76.0$	$\Sigma y^2 = 69\,864$	

Data obtained from MBL Group PTA00292

From Table 13.10 it follows that the least squares values of a and b are given by

$$b = \frac{111\,632 - \frac{1}{12}(1\,476 \times 912)}{182\,120 - \frac{1}{12}(1\,476)^2}$$

$$= -0.951049$$

$$a = 76.0 - (-0.9510 \times 123.0)$$

$$= 192.9790$$

The regression line of y given x is therefore given by

$$\hat{y} = 192.98 - 0.9510x$$

Now suppose we wish to determine a 95% confidence band for the population regression line $\mu_y = \alpha + \beta x$ for the interval of x values from 112 up to and including 134 knots. Table 13.11 shows the required calculations for obtaining the confidence intervals for $\alpha + \beta x$ based on the selected x values within this interval. In calculating these intervals the value $t_{10,\,0.025} = 2.228$ and the sample standard deviation around the regression line $\hat{y} = 192.98 - 0.9510x$, namely

$$s_{y \cdot x} = \sqrt{\tfrac{1}{10}\{69\,864 - 192.979021 \times 912 - (-0.951049 \times 111\,632)\}}$$

$$= 1.861$$

Table 13.11 95% confidence intervals necessary for compiling a 95% confidence band for the regression line $\alpha + \beta x$

x	$\hat{y}$	$(x - \bar{x})^2$	$d_x = s_{y \cdot x}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}}$	$\hat{y} \pm 2.228 \times d_x$
112	86.467	121	1.0100	(84.22, 88.72)
114	84.565	81	0.8820	(82.60, 86.53)
116	82.663	49	0.7643	(80.96, 84.37)
118	80.761	25	0.6624	(79.29, 82.24)
120	78.859	9	0.5848	(77.56, 80.16)
122	76.957	1	0.5418	(75.75, 78.16)
124	75.055	1	0.5418	(73.85, 76.26)
126	73.153	9	0.5848	(71.85, 74.46)
128	71.251	25	0.6624	(69.78, 72.73)
130	69.349	49	0.7643	(67.65, 71.05)
132	67.447	81	0.8820	(65.48, 69.41)
134	65.545	121	1.0100	(63.29, 67.80)

The confidence intervals are represented in Figure 13.3 together with the regression line of y given x, namely $\hat{y} = 192.98 - 0.951x$. If the calculated upper interval values, namely $\hat{y} + 2.228d_x$ are joined and the same is done to the calculated lower interval values, namely $\hat{y} - 2.228d_x$, the 95% confidence band is obtained for $\alpha + \beta x$. Note that this band is narrowest where $x = \bar{x}$, in other words where $x = 123$, and broadens as the difference between x and $\bar{x}$ increases.

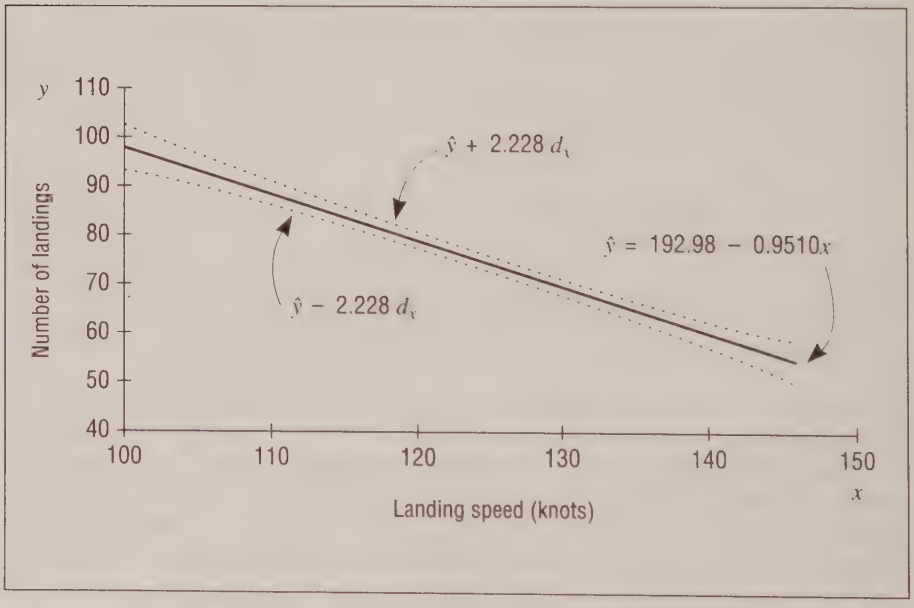


Figure 13.3 Confidence band for $\mu_y = \alpha + \beta x$ based on the bivariate sample of size 12 given in Table 13.10

6 Prediction

We are often interested in making a prediction, based on the information of an existing sample, of an individual value, for example y_0, associated with a particular x value, x_0. The flight engineer, for instance, wants to estimate the number of landings that an aircraft will make with a particular set of tyres if the landing speed is 140 knots. This prediction of the number of landings, Y_0, can be made by means of a 95% confidence interval. When determining a confidence interval for an individual y value x_0 is usually chosen within the interval of currently observed x values. The reason for this is that the linear relationship does not necessarily remain valid outside this interval. Also, the 95% confidence interval for an individual y_0 value increases in length as x_0 deviates from $\bar{x}$. However, in practice we often find in, for example, time series analysis, that we are in fact interested in a prediction outside the area of observed x values. This kind of prediction should be approached cautiously.

Under the inference assumption regarding linear regression and based on the observed sample $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ the **95% confidence interval for the individual value**, y_0, associated with x_0, is given by

$$(a + bx_0 - t_{n-2,\,0.025}\, d_0,\ a + bx_0 + t_{n-2,\,0.025}\, d_0)$$

with d_0 the observed standard deviation of Y_0, namely

$$d_0 = s_{y \cdot x} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}}$$

The graphic representation of the 95% confidence intervals for y_0, based on a set of given x values, leads to the so-called **95% prediction band** (see Figure 13.4).

EXAMPLE 13.8

To illustrate the calculation of a 95% prediction band we again use the data in Example 13.7, Table 13.10. To compare, for given x values, the confidence intervals for individual y values with the confidence intervals for the regression line $\mu_y = \alpha + \beta x$, the landing speed values in Table 13.10 are used. Table 13.12 shows the calculations. Remember that $t_{10,\,0.025} = 2.228$.

Figure 13.4 is a graphic representation of both the 95% prediction and confidence bands for the data in Table 13.10. This figure clearly shows that, throughout, the prediction band is substantially broader than the confidence band. The difference between the two bands can be explained as follows in terms of the problem under consideration: The confidence interval for $\alpha + \beta x$ is used to estimate the average number of landings, μ_y, on a set of tyres at a given landing speed, while the confidence interval for the individual value y_0 is used to estimate the number of landings by a particular aircraft with a set of tyres at a given landing speed.

Table 13.12 95% confidence intervals for individual y values needed to compile a 95% prediction band

x_0	$\hat{y}_0 = a + bx_0$	$(x_0 - \bar{x})^2$	$d_0 = s_{y\,x}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}}$	$\hat{y}_0 \pm 2.228 d_0$
112	86.47	121	2.1174	(81.75, 91.19)
114	84.57	81	2.0594	(79.98, 89.15)
116	82.66	49	2.0118	(78.18, 87.15)
118	80.76	25	1.9754	(76.36, 85.16)
120	78.86	9	1.9507	(74.51, 83.21)
122	76.96	1	1.9383	(72.64, 81.28)
124	75.06	1	1.9383	(70.74, 79.37)
126	73.15	9	1.9507	(68.81, 77.50)
128	71.25	25	1.9754	(66.85, 75.65)
130	69.35	49	2.0118	(64.87, 73.83)
132	67.45	81	2.0594	(62.86, 72.04)
134	65.55	121	2.1174	(60.83, 70.26)

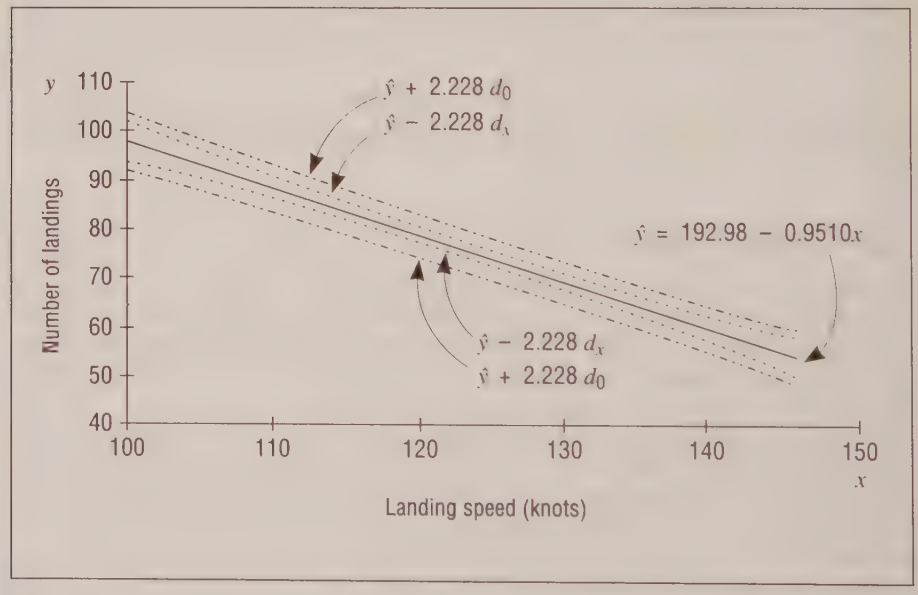


Figure 13.4 Comparison of the 95% confidence and prediction bands based on a bivariate sample of size 12 given in Table 13.10

7 Inference regarding the correlation coefficient ρ

In Section 1 we demonstrated in the observation of two population characteristics that r, the sample correlation coefficient, varies from one sample to the next and that the different r values can be regarded as outcomes of the stochastic variable R. R is an estimator of the unknown population correlation coefficient ρ which is a measure of the strength of the linear relationship that exists in the population between the two characteristics X and Y. Figure 13.5 illustrates a number of possible ρ values by means of scatter diagrams.

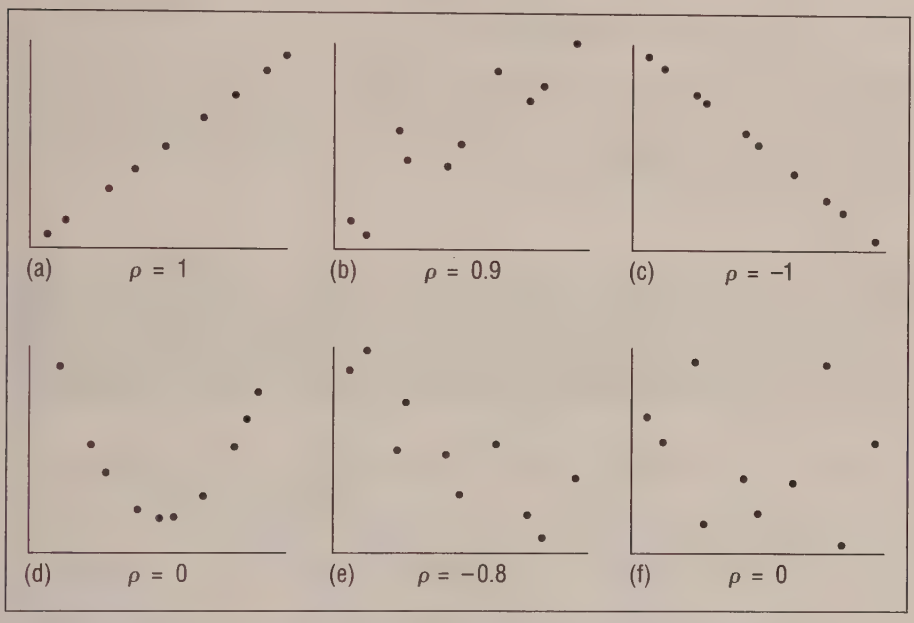


Figure 13.5 Scatter diagrams of bivariate samples of size 10 from populations with different values of ρ

It can be proved that $-1 \leq \rho \leq 1$ where $\rho = 1$ indicates a perfect positive and $\rho = -1$ a perfect negative linear relationship between X and Y (see Figure 13.5 (a) and (c)). Since ρ is only a measure of linear association, it may occur that $\rho = 0$ despite the fact that there is a perfect non-linear relationship between X and Y (see Figure 13.5(d)). In this section the statistical inference regarding the parameter ρ is discussed and for this the following assumption is required:

Assumption for inference regarding linear correlation: The bivariate data set $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ is an observed random sample from a population with characteristics X and Y. These stochastic variables have a so-called bivariate distribution and the correlation coefficient between X and Y is ρ.

7.1 Testing H_0: $\rho = 0$

The following result is important when testing the above hypothesis:

> Suppose the inference assumption regarding linear correlation is met. If X and Y are both normally distributed stochastic variables and if $\rho = 0$, the sample correlation coefficient R has the so-called **r(n) distribution**.

7.1.1 Tables for the r(n) distribution

Since the r(n) distribution is symmetrical around 0, it will suffice to tabulate only positive critical values of this distribution. Figure 13.6 illustrates the population curve of the r(20) distribution. If R_n has the r(n) distribution, the value for $r_{n,\varepsilon}$ is defined by

$$P(R_n \geq r_{n,\varepsilon}) = \varepsilon$$

Table B.8 of Appendix B gives the values of $r_{n,\varepsilon}$ for $n \geq 3$ and selected values of ε. The starting point in the table is at $n = 3$ since the r(n) distribution is not defined for $n = 1$ and $n = 2$. The above results can now be used to test H_0 against the appropriate alternative. Table 13.13 gives a summary of the test procedure.

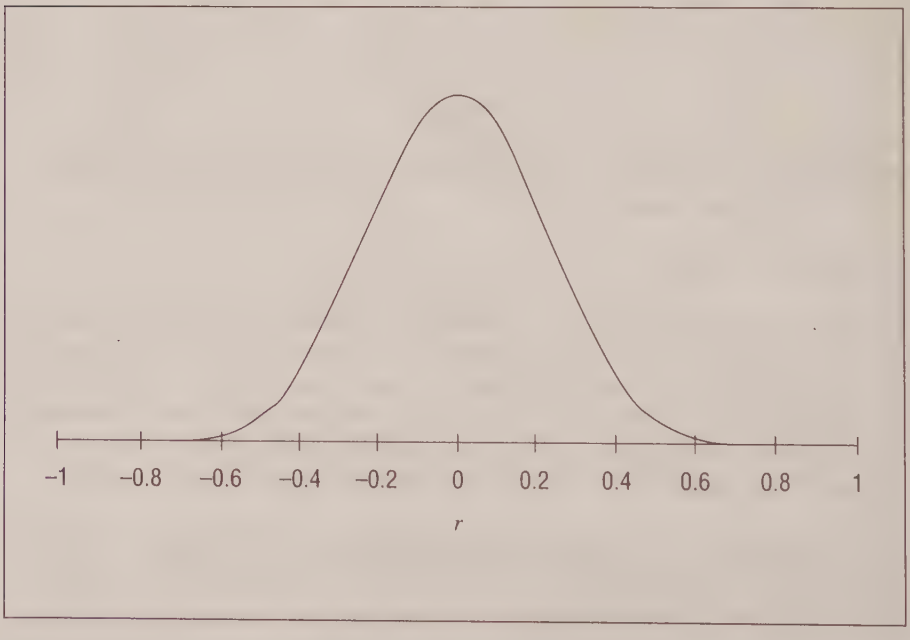


Figure 13.6 Population curve of the sample distribution of R, based on 20 observations, under the assumption that $\rho = 0$

Table 13.13 Testing H_0: $\rho = 0$, based on a bivariate random sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ where both X and Y have normal distributions

Test statistic	Distribution of test statistic under H_0	Table with critical values
R	$r(n)$	Tabel B.8
Alternative hypothesis	**Decision rule: Reject H_0 if**	**p-value**
$H_1: \rho \neq 0$	$r \leq -r_{n,\varepsilon/2}$ or $r \geq r_{n,\varepsilon/2}$	$2P(R \geq \lvert r \rvert)$
$H_1: \rho > 0$	$r \geq r_{n,\varepsilon}$	$P(R \geq r)$
$H_1: \rho < 0$	$r \leq -r_{n,\varepsilon}$	$P(R \leq r)$

EXAMPLE 13.9

In the television series *Beyond 2000* it was alleged that there is a negative linear correlation between income and the number of hours that a person sets aside for sleep. To investigate this statement, a random sample of twenty people, all working for the same company, was taken. The following data was obtained:

Table 13.14 Monthly income and average number of hours slept per day by twenty employees

Observation	Numbers of hours slept x	Monthly income (R1 000) y
1	7.0	0.844
2	9.3	1.708
3	7.4	1.728
4	8.9	2.909
5	7.7	2.676
6	8.9	3.440
7	7.3	3.616
8	6.3	4.096
9	7.7	4.420
10	7.6	5.808
11	8.1	5.440
12	7.5	6.065
13	7.4	6.444
14	7.3	7.236
15	7.6	7.349
16	7.1	8.235
17	7.4	8.379
18	5.9	9.131
19	6.0	9.746
20	6.7	9.983

Suppose we wish to determine whether or not there is a negative linear correlation between X and Y. We shall go about it as follows:

Step 1

H_0: $\rho = 0$
H_1: $\rho < 0$

Step 2

Let $\varepsilon = 0.05$. If H_0 is true and the income and hours slept X and Y are normally distributed, the sample correlation coefficient R has an r(20) distribution.

Step 3

Since it follows from Table B.8 that $r_{20,\,0.05} = 0.37834$, H_0 will be rejected if $r \leq -0.37834$.

Step 4

Calculations yield $\Sigma x = 149.1$, $\Sigma y = 109.253$, $\Sigma x^2 = 1\,126.53$, $\Sigma y^2 = 747.507667$ and $\Sigma xy = 787.6877$ so that

$$r = \frac{\sum xy - \frac{1}{n}\sum x \sum y}{\sqrt{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}\sqrt{\sum y^2 - \frac{1}{n}\left(\sum y\right)^2}}$$

$$= \frac{787.6877 - \frac{1}{20}149.1 \times 109.253}{\sqrt{1\,126.53 - \frac{1}{20}(149.1)^2}\sqrt{747.507667 - \frac{1}{20}(109.253)^2}}$$

$$= -0.5637$$

Step 5

Since $r < -0.37834$, it follows that H_0 is rejected at a 5% level of significance. Income and number of hours slept therefore do show a negative linear correlation.

Notes

(a) An examination of Table B.8 clearly shows that the critical values for testing H_0: $\rho = 0$ decrease as n increases. Suppose, for example, that $n = 50$ and $r = -0.30$, then H_0: $\rho = 0$ is rejected at a 5% level of significance in favour of H_1: $\rho \neq 0$ and the conclusion is that ρ differs significantly from 0. However, in Chapter 6 we see that r^2 indicates the percentage variation in y which is explained by x. For $r = -0.30$ it therefore follows that $r^2 = 0.09$, so that only 9% of the variation in the y values is explained by a linear relationship with the x values. A significant correlation coefficient is therefore not necessarily an indication of a strong

linear relationship. An absolute correlation coefficient of 0.7 and greater implies that at least 50% of the variation in y is explained by a linear relationship with the x values. In view of this consideration a general guideline is that $|r| \geq 0.7$ is an indication of a strong linear relationship (if n is accepted as being sufficiently large enough for H_0: $\rho = 0$ to be rejected).

(b) Since the r(n) distribution has been defined only for $\rho = 0$, it cannot be used to test H_0: $\rho = \rho_0$ or to compile a confidence interval for ρ.

7.2 Confidence interval for ρ (the z transformation)

In view of Note (a) above it is often more important to obtain a confidence interval for ρ than to test H_0: $\rho = 0$. As already stated, the r(n) distribution cannot be applied to obtain such a confidence interval, but it is possible to find an approximate 95% confidence interval for ρ by means of the so-called z **transformation of** r. For this, the following result is required:

If n is large, then

$$z(r) = \tfrac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

is, under the assumption for inference regarding linear correlation, an observation of a stochastic variable which approximately has the normal distribution with a mean of

$$z(\rho) = \tfrac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$

and a variance of $\frac{1}{n-3}$.

From the above result it follows that the **pivotal quantity**

$$Z = \{z(R) - z(\rho)\}\sqrt{n-3}$$

is a stochastic variable which is approximately distributed as n(0,1). Table B.9 was compiled to simplify the calculations. For a wide variety of values of r the corresponding values of $z(r)$ can be read directly from this table. Note that, since $z(-r) = -z(r)$, the table has been compiled only for positive values of z. The above pivotal quantity now leads to an approximate 95% confidence interval for $z(\rho)$.

If n is large and the inference assumption regarding linear correlation is met, an **approximate 95% confidence interval for** $z(\rho)$ is given by

$$\left(z(r) - \frac{1.96}{\sqrt{n-3}},\ z(r) + \frac{1.96}{\sqrt{n-3}}\right)$$

Using Table B.9 the lower and upper bounds of this interval can be used to obtain an approximate 95% confidence interval for ρ. This technique is illustrated in Example 13.10.

Note

The upper and lower bounds of the confidence interval (ρ_1, ρ_2) for the correlation coefficient can also be calculated directly with $\rho_1 = \frac{e^{2z_1} - 1}{e^{2z_1} + 1}$ and $\rho_2 = \frac{e^{2z_2} - 1}{e^{2z_2} + 1}$ and where z_1 and z_2 are the upper and lower bounds of the confidence interval for $z(\rho)$.

EXAMPLE 13.10

In an investigation into the relationship between the floor area, x, and the selling price, y, in Cape Town the information in Appendix A, Data Set 2, was used. Suppose we use the data on houses 1 to 40 to compile an approximate 95% confidence interval for ρ. Routine calculations give

$\Sigma x = 6\,001$, $\Sigma y = 4\,861.3$, $\Sigma x^2 = 1\,049\,731$, $\Sigma y^2 = 694\,520.75$ and $\Sigma xy = 802\,532.80$, and that

$$r = \frac{\sum xy - \frac{1}{n}\sum x \sum y}{\sqrt{\sum x^2 - \frac{1}{n}(\sum x)^2}\sqrt{\sum y^2 - \frac{1}{n}(\sum y)^2}}$$

$$= \frac{802\,532.80 - \frac{1}{40} \times 6\,001 \times 4\,861.3}{\sqrt{1\,049\,731 - \frac{1}{40}(6\,001)^2}\sqrt{694\,520.75 - \frac{1}{40}(4\,861.3)^2}}$$

$$= 0.5881 \approx 0.59$$

From Table B.9 it follows that $z(0.59) = 0.67767$ so that an approximate 95% confidence interval for $z(\rho)$ is given by

$$\left(0.67767 - \frac{1.96}{\sqrt{37}},\ 0.67767 + \frac{1.96}{\sqrt{37}}\right),$$

which is (0.3554, 0.9999).

The corresponding lower and upper bounds for ρ are now found as follows: Search in Table B.9 for the figure closest to the lower z value 0.3554. In this case this figure is 0.35409. Now move to the left in this row and upwards into the relevant column in order to read the matching r value, namely 0.34. Similarly, it will be found that $r = 0.76$ corresponds approximately with the upper z value of 0.9999. An approximate 95% confidence interval for ρ is therefore (0.34, 0.76). With the aid of the transformation given in the foregoing remark, the interval (ρ_1, ρ_2) is calculated as follows:

$\rho_1 = \frac{e^{2\times0.3554} + 1}{e^{2\times0.3554} - 1} = 0.341$ and $\rho_2 = \frac{e^{2\times0.9999} + 1}{e^{2\times0.9999} - 1} = 0.762$, so that an approximate confidence interval for ρ is the interval (0.34, 0.76).

7.3 Testing H_0: $\rho_1 = \rho_2$

Consider two independent bivariate samples of size n_1 and n_2, which have been taken respectively from populations with correlation coefficients ρ_1 and ρ_2. We assume that both data sets comply with the assumption of inference regarding linear correlation. If n_1 and n_2 are large, the test procedure for H_0: $\rho_1 = \rho_2$, as set out in Table 13.15, is used.

Table 13.15 Testing H_0: $\rho_1 = \rho_2$ based on two independent bivariate samples of sizes n_1 and n_2 respectively

Test statistic	Distribution of test statistic under H_0	Table with critical values
$Z = \dfrac{Z(R_1) - Z(R_2)}{\sqrt{\dfrac{1}{n_1 - 3} + \dfrac{1}{n_2 - 3}}}$	n_1 n_2 large; approximately n(0, 1)	Tabel B.2
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
H_1: $\rho_1 \neq \rho_2$ or $\rho_1 - \rho_2 \neq 0$	$z \leq -z_{\varepsilon/2}$ or $z \geq z_{\varepsilon/2}$	$2P(Z \geq \lvert z \rvert)$
H_1: $\rho_1 > \rho_2$ or $\rho_1 - \rho_2 > 0$	$z \geq z_{\varepsilon}$	$P(Z \geq z)$
H_1: $\rho_1 < \rho_2$ or $\rho_1 - \rho_2 < 0$	$z \leq -z_{\varepsilon}$	$P(Z \leq z)$

EXAMPLE 13.11

It may be important to test whether or not the correlation coefficients ρ_1 and ρ_2 differ for the floor area, x, and the selling price, y, in the Cape and Johannesburg respectively. In Example 13.10 we find that $r_1 = 0.59$ with $n_1 = 40$. To test H_0: $\rho_1 = \rho_2$, we now consider the data of the first 40 Johannesburg houses in Appendix A. The test procedure is as follows:

Step 1

H_0: $\rho_1 = \rho_2$
H_1: $\rho_1 \neq \rho_2$

Step 2

Let $\varepsilon = 0.05$. Under H_0, $z = \dfrac{z(r_1) - z(r_2)}{\sqrt{\frac{1}{37} + \frac{1}{37}}}$ can be regarded as an observation from a n(0, 1) distribution.

Step 3

Reject H_0 if $z \leq -1.96$ or $z \geq 1.96$ or if the p-value ≤ 0.05.

Step 4

From the data in Appendix A it follows, for the Johannesburg houses, that $n_2 = 40$, $\Sigma x = 7\,904$, $\Sigma y = 7\,416.40$, $\Sigma x^2 = 1\,664\,584.00$, $\Sigma y^2 = 1\,616\,139.66$ and $\Sigma xy = 1\,575\,011.20$. From this it follows, after calculations, that $r_2 = 0.70$. With the aid of Table B.9 it is now found that $z(0.59) = 0.67767$ and

$z(0.70) = 0.86730$ or, by means of transformation

$$z(0.59) = \tfrac{1}{2}\ln\left(\tfrac{1+r}{1-r}\right) = \tfrac{1}{2}\ln\left(\tfrac{1.59}{0.41}\right) = 0.67767 \text{ and}$$

$$z(0.70) = \tfrac{1}{2}\ln\left(\tfrac{1.70}{0.30}\right) = 0.86730 \text{ so that}$$

$$z = \frac{0.67767 - 0.86730}{\sqrt{\frac{1}{37} + \frac{1}{37}}} = -0.82$$

Step 5

Since $-1.96 < z < 1.96$, H_0 is not rejected. There is therefore not sufficient reason to conclude that ρ_1 and ρ_2 differ. The p-value is

$$2P(Z \geq |-0.82|) = 2[0.5 - 0.29389] = 0.41222.$$

8 The trivariate linear model

Suppose there is a linear relationship between the dependent variable y and the two explanatory variables x and z. As already mentioned in Chapter 6, Section 9, this relationship can, in the case of a sample of trivariate data, be represented by the following equation:

$$\hat{y} = a + b_1x + b_2z$$

The sample coefficient of determination, namely

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

indicates how well the plane $\hat{y}$ describes the observed data set. Although, in Chapter 6, we used a capital letter to distinguish the coefficient of determination from the observed correlation coefficient r, R^2 is an observed value of a stochastic variable for every sample. The statistical inference regarding the **population regression plane of y given x and z** is based on the following important result:

> Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent stochastic variables that correspond respectively with the given $(x_1, z_1), (x_2, z_2), \ldots, (x_n, z_n)$.
>
> Also suppose that Y_i has a $n(\alpha + \beta_1 x_i + \beta_2 z_i, \sigma^2)$ distribution, $i = 1,2, \ldots, n$, then it follows that
>
> $$f = \frac{(n-3)R^2}{2(1-R^2)}$$
>
> is an observation of an $F(2, n-3)$ distribution, provided that the coefficient of determination of the population is zero.

This result can be used to test the following two equivalent null hypotheses:

(a) H_0: The coefficient of determination of the population is zero.

(b) H_0: $\beta_1 = \beta_2 = 0$

In reality both the above hypotheses imply that there is no linear relationship between y as a dependent variable and x and z as jointly explanatory variables. This means that the population regression plane of y given x and z, namely $\mu_y = \alpha + \beta_1 x + \beta_2 z$, reduces to $\mu_y = \alpha$. Note that the rejection of the null hypothesis really means that the coefficient of determination of the population is larger than zero.

EXAMPLE 13.12

In Example 6.15 we set out to see whether or not there is a linear relationship between building valuation, y, and the explanatory variables of land valuation, x, and the total floor area, z. For a random sample of ten houses it was found that $R^2 = 0.81$. Suppose it is necessary to test whether the population coefficient of determination is larger than zero. The testing procedure is as follows:

Step 1

H_0: The coefficient of determination of the population is zero.
H_1: The coefficient of determination of the population is larger than zero.

Step 2

Let $\varepsilon = 0.05$. Under H_0, $f = \dfrac{7 \times R^2}{2 \times (1 - R^2)}$ is an observation from an F(2,7) distribution if we assume that the building valuation is taken from a normal population and has the same variance for every given (x, z).

Step 3

Reject H_0 if $f \geq F_{2,7;0.05} = 4.74$ (Table B.5).

Step 4

For $R^2 = 0.81$, it follows that $f = \dfrac{7 \times 0.81}{2 \times (1 - 0.81)} = 14.921$

Step 5

Since $f > 4.74$ it follows that H_0 is rejected at a 5% level of significance. Table B.5 shows that $F_{2,7;0.01} = 9.55$ so that H_0 is also rejected at a 1% level of significance. Building valuation can therefore be described linearly in terms of land valuation and total floor area.

9 Concluding remarks

9.1 The general linear model

In this chapter statistical inference is considered only in situations where there is a linear relation between the dependent variable and one or two explanatory variables. The inference techniques described in this chapter can, however, be extended directly to the case where there is a linear relation between the dependent variable and more than two explanatory variables. The equation describing such a relationship is known as the **general linear model**.

9.2 Calculation aids

There is a wide variety of computer programmes that execute linear regression analyses for bivariate, trivariate or multivariate data sets. Moreover many pocket calculators can now be used to compute values for a, b and r for fitting the straight line $\hat{y} = a + bx$ by simply keying in the observed bivariate data set.

Exercises

In all of the following exercises you can accept that, where applicable, the inference assumptions regarding linear regression and correlation are valid.

1 In order to investigate the relationship between valuation, x, and the selling price, y, of houses in Cape Town, the information for houses with numbers 101 to 110 (Appendix A, Data Set 1, columns 4 and 3 respectively) is used. Determine the regression line of y, given x, and then test the following statements at a 5% level of significance.

(a) $\alpha = 0$ (the population intercept is zero);

(b) $\beta = 0.9$ (the population slope is 0.9).

2 Repeat Exercise 1 for the Johannesburg houses with numbers 501 to 510.

3 Use the data in sample 3 of Example 13.1 (see Tables 13.1 to 13.3) to

(a) calculate a 99% confidence interval for α;

(b) calculate a 99% confidence interval for β.

4 Use the data in sample 5 of Example 13.1 (also see Tables 13.1 to 13.3) to

(a) calculate a 99% confidence interval for α;

(b) calculate a 99% confidence interval for β.

5 Production in a particular fertilizer factory (in hundred metric tons) for the period 1987 to 1993 was as follows:

Year	1987	1988	1989	1990	1991	1992	1993
Production, y	28.6	33.6	36.2	50.8	53.1	49.7	56.2

(a) Determine $\hat{y} = a + bx$ by transforming the dates to x values in such a way that 1990 and $x = 0$ correspond.

(b) Find a 95% confidence interval for $\alpha + \beta x$, where $x = 5$, in other words, a confidence interval for the expected 1995 production.

6 The following data shows the age (in years) and the selling price (R1 000) of used cars with the same engine capacity and make at eight different garages.

Age, x	1	6	4	2	5	4	1	2
Price, y	41.250	10.250	24.310	38.720	8.740	26.110	38.650	36.200

(a) Calculate the equation of the regression line of y given x.

(b) Find a 95% confidence interval for the individual value y_0 that corresponds with $x_0 = 4$, in other words a confidence interval for the price for which a single four-year-old car (of the kind mentioned above) may possibly be sold.

7 The following data shows the relationship between the number of old age homes and the population size (in thousands) as observed in twelve cities.

Number of old age homes, y	13	17	31	46	9	70	17	23	10	13	21	38
Population size, x	112	145	283	443	84	654	157	204	96	124	189	362

(a) Determine $\hat{y} = a + bx$.

(b) Calculate a 95% confidence interval for $\alpha + \beta x$ if $x = 200$, in other words, for the expected number of old age homes in a city with a population of 200 000.

8 Test the statement that an increase of 10 000 in population size justifies more than one additional old age home (see Exercise 7) and use a significance level of 0.05.

9 Various researchers have tried to establish the minimum number of hours of sleep required daily by a child of a particular age (in years). The following data is based on the findings of a group of researchers:

Age, x	1.5	2.0	3.5	4.5	5.5	7.5	13.5	17.0
Minimum amount of sleep required, y	15.5	14.5	13.0	12.0	11.5	10.5	8.5	8.0

(a) Calculate the sample correlation coefficient, r.

(b) Test for a negative linear dependency between the ages and the hours of sleep required daily and use a significance level of 0.01.

10 The average number of defective parts (per hour) is believed to be directly proportional to the speed (in revolutions per minute) of the machine producing this part. Nine working days selected at random in a month, yield the following data:

Speed of machine, x	129	150	82	168	133	112	106	186	149
Number of defective parts, y	9.3	12.2	6.1	14.3	9.6	7.5	5.7	12.3	9.2

(a) Determine $\hat{y} = a + bx$.

(b) Does a have any practical significance in this situation?

(c) Calculate a 95% confidence interval for β.

11 Calculate the sample correlation coefficient between the valuation, x, and the selling price, y, of residences in the Cape with numbers 101 to 110 (see Exercise 1). Also determine r for the sample of Johannesburg houses with numbers 501 to 510 (Exercise 2) and then test

H_0: $\rho_1 = \rho_2$ against H_1: $\rho_1 \neq \rho_2$ at a significance level of 1%.

12 Two independent random samples yield the following data:

Sample 1	Sample 2
$n_1 = 65$	$n_2 = 75$
$r_1 = 0.83$	$r_2 = 0.78$

Use this information to test H_0: $\rho_1 = \rho_2$ against H_1: $\rho_1 > \rho_2$ (1% level of significance).

13 Use the data in sample 3 of Example 13.1 (see also Exercise 3) to test for positive linear correlation between valuation and selling price.

14 Use the data in sample 5 of Example 13.1 to test for positive linear correlation between valuation and selling price.

15 Calculate a 95% confidence interval for ρ if

(a) $n = 34$, $r = 0.87$;

(b) $n = 87$, $r = 0.72$;

(c) $n = 221$, $r = 0.23$;

(d) $n = 10$, $r = 0.95$;

(e) $n = 100$, $r = 0.30$;

(f) $n = 500$, $r = 0.18$.

16 Compile a 95% prediction band for individual y_0 values based on the data in Exercise 1 (see Example 13.8).

17 Compile a 99% confidence band for the regression line $\mu_y = \alpha + \beta x$ by using the data in Exercise 6 (see Example 13.7).

18 Compile a 95% prediction band for individual y_0 values based on the data in Exercise 9 (see Example 13.8).

19 Compile a 95% prediction band for individual y_0 values based on the data in Exercise 10 (see Example 13.8).

20 Refer to Chapter 6, Exercise 12. Test H_0: $\rho = 0$ against H_1: $\rho \neq 0$. What conclusion can you reach concerning the works engineer's statement? (Level of significance = 0.05.)

21 (a) Refer to Chapter 6, Exercise 18. Test H_0: $\beta = 1$ against H_1: $\beta > 1$.

(b) Calculate a 95% confidence interval for the individual value y_0 if $x_0 = 3$.

22 (a) Refer to Chapter 6, Exercise 19. Test H_0: $\alpha = 20$ against H_1: $\alpha \neq 20$.

(b) Compile a 95% confidence interval for y if $x = 25$.

23 (a) Refer to Chapter 6, Exercise 20. Compile a 95% confidence interval for

(i) α;
(ii) β.

(b) Test H_0: $\rho = 0$ against H_1: $\rho > 0$ with a significance level of 0.05.

24 Refer to the data of Chapter 6, Exercise 22. Test H_0: $R^2 = 0$ against H_1: $R^2 \neq 0$ (level of significance = 0.05).

25 Repeat Example 13.12 (see also Example 6.13) for the last ten houses in Durban.

26 (a) Fit a straight line to the data in Chapter 6, Exercise 24.
(b) Compile a 95% interval for y if $x = 1995$.

27 (a) Fit a straight line to the data in Chapter 6, Exercise 26.
(b) Compile a 95% confidence interval for the individual value y_0 if $x_0 = 1990$.

28 The management of an airline suspects that there is a direct link between advertising expenditure and the number of passengers who prefer to fly with the airline. The expenditure on advertisements, and the number of passengers for the past twelve months are used and the data appears in the accompanying table.

Advertising expenditure (R1 000)	Passengers (1 000)
11.2	16.8
13.3	19.0
8.5	12.8
16.9	24.3
11.3	16.4
15.9	21.8
11.0	15.4
15.1	21.7
20.3	25.3
10.5	18.0
12.6	16.6
14.0	19.5

(a) Fit a linear regression equation to predict the number of passengers from advertising expenditure.

(b) How many passengers can be expected for the following month if the budgeted advertising expenditure amounts to R19 900? Give a 95% confidence interval for your estimate.

(c) Test whether the regression coefficient for the number of passengers, given expenditure, is equal to 1.1.

(d) Interpret the regression coefficient determined in (a).

29 Use the data in Appendix A, Data Set 5. Let y be the weekly amount of electricity consumed (kWh) and x the weekly amount of beer produced.

(a) Fit a linear regression equation to predict the amount of electricity consumed and the amount of beer produced in each of the two years being reviewed.

(b) Calculate the correlation coefficients between the amount of electricity consumed and the amount of beer produced for both years under review.

(c) Test H_0: $\rho_1 = \rho_2$ against H_1: $\rho_1 < \rho_2$ with a level of significance = 0.05.

(d) What does a, the estimated intercept value, indicate in both cases?

(e) Test the hypothesis that the intercept is equal to 100 000 kWh for both the 1992 and 1993 fiscal years. Use a level of significance of 0.01.

(f) Test the hypothesis that the slope is equal to 5 for both years. Use a level of significance of 0.01.

(g) Interpret the slope of the regression line for each of the years.

(h) Calculate a 95% confidence interval for the intercept for both years.

(i) Calculate a 95% confidence interval for the slope for both years.

(j) Write a summary of your findings and decide whether the project that was launched to cut electricity consumption in the 1993 fiscal year was a success (see also the results of Chapter 12, Exercises 59, 60 and 61).

30 In a feedlot experiment for oxen, the relation between the cumulative amount of feed (kg) consumed was measured every week. The data for two of the oxen is as follows:

Feeding days x	**Cumulative amount of feed (kg)**	
	Ox 1	Ox 2
0	0	0
8	15	40
15	71	105
22	132	170
29	199	236
36	269	307
43	328	369
50	391	432
57	455	498
64	531	575
71	570	620
78	682	733

(a) Fit a linear regression equation of cumulative amount of feed consumed against feeding days for each ox with y = cumulative amount of feed and x = number of feeding days.

(b) Calculate the correlation coefficients between the cumulative amount of feed and the feeding days for both oxen.

(c) Test H_0: $\rho_1 = \rho_2$ against H_1: $\rho_1 \neq \rho_2$ with a level of significance = 0.05, where ρ_1 is the correlation coefficient for ox 1 and ρ_2 is the correlation coefficient for ox 2.

(d) Estimate the cumulative consumption of ox 1 on 40 and on 80 feeding days. Give a 95% confidence interval for your estimate.

31 Consider Chapter 12, Exercise 58.

(a) Fit a linear regression equation by taking the old and the new method as x and y variables respectively.

(b) Calculate the correlation coefficient and calculate a 95% confidence interval for ρ.

(c) Test the hypothesis that the population intercept is equal to zero with a level of significance = 0.05.

(d) Test the hypothesis that the population slope is equal to 1 with a level of significance = 0.05.

(e) Can the old method be replaced by the new one? Substantiate your answer by discussing the results of questions (b), (c) and (d).

32 Consider Chapter 6, Exercise 32.

(a) Fit a linear regression equation to the data.

(b) Estimate the petrol price if the mean oil price is \$15 per barrel and give a 95% confidence interval for your estimate.

(c) Estimate the petrol price if the mean oil price is \$25 per barrel and give a 95% confidence interval for your estimate.

(d) Set a 95% prediction band for individual y_0 values.

(e) Calculate the correlation between petrol price and oil price and give a 95% confidence interval for ρ.

33 Consider the data in Chapter 6, Exercise 32.

(a) Draw a scatter diagram of oil price against time. Use as x values the numbers 1, 2, ..., 36 to correspond with January 1988 to December 1990.

(b) Fit a straight line to the data.

(c) Estimate the oil price for January 1991 and for January 1995.

(d) Determine a 95% prediction band for individual y_0 values.

34 Consider the data in Chapter 6, Exercise 32, and test H_0: $R^2 = 0$ against H_1: $R^2 \neq 0$ (level of significance = 0.05).

35 Consider the turnover and the number of clients in Appendix A, Data Set 4, for the month of April.

(a) Fit a straight line to the data.

(b) Estimate the turnover if there are 100 customers in the restaurant and give a 95% confidence interval for your estimate.

(c) Calculate the correlation between the turnover and the number of customers and test whether the correlation is equal to 0 with a level of significance equal to 0.01.

36 Repeat Exercise 35 for the month of May.

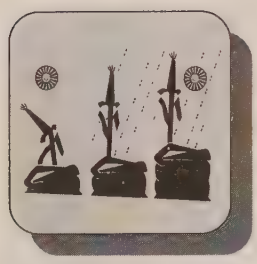
37 A marketing manager hired a consultant to develop a multiple regression model to predict sales, Y (R1 000) by using advertising expenditure X (R1 000) and market share Z (as a percentage).

Sales y	Advertising expenditure x	Percentage market share z
10.8	4.3	26.2
12.6	4.0	32.2
8.3	4.1	17.3
9.2	4.6	16.7
11.1	5.5	18.9
10.9	4.5	13.2
7.9	3.3	14.4
11.6	2.8	27.1
8.2	2.6	20.8
9.0	3.1	22.0
13.3	4.3	33.3
7.2	2.6	15.0

(a) Fit a curve of the form $\hat{y} = a + bx + cz$ to the data.

(b) Calculate the estimated sales $\hat{y}$ for each observation.

(c) Calculate R^2 and test H_0: $R^2 = 0$ against H_1: $R^2 \neq 0$. Use a significance level of 0.05.



CHAPTER 14

Analysis of variance

1 Introduction

Testing hypotheses regarding the parameters of one or two populations were discussed in Chapter 12. However, it is often very important to test whether various populations differ from one another in respect of a particular characteristic. For the agricultural economist it is important to know whether different wheat varieties differ from one another in respect of the yield per hectare. For the industrialist, on the other hand, it is vitally important to determine the differences in the productivity of workers who have undergone three diverse training programmes. A pharmaceutical company will, for instance, market a new remedy only if it compares well, as far as effectivity is concerned, with all the existing remedies for the same disease. The statistical analysis technique, known as **analysis of variance**, can be used to provide solutions in these situations. This technique is used to test whether the means of a number of populations differ from one another.

Suppose $\mu_1, \mu_2, \ldots, \mu_c$ are the means of c populations. If a random sample is available from each of these populations and certain assumptions are met, analysis of variance involves the following hypothesis-testing procedure

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_c$$

against

H_1: At least one of the c means differs from the others.

Analysis of variance is based on the division of the total variance of all c samples into different variance components. A mutual comparison of all of these components leads to a test for H_0. The designation of this important analysis technique therefore refers to the **test statistic** and not to the hypothesis being tested.

Testing $H_0: \mu_1 = \mu_2$ was discussed in Chapter 12. Should we now wish to test the equality of the means of, for example, three populations, namely

$$H_0: \mu_1 = \mu_2 = \mu_3,$$

this can be done by testing

H_0: $\mu_1 = \mu_2$, H_0: $\mu_1 = \mu_3$ and H_0: $\mu_2 = \mu_3$

successively. If none of these three hypotheses are rejected, it means that H_0: $\mu_1 = \mu_2 = \mu_3$ is apparently not rejected. If, however, one or more of the latter series of null hypotheses is rejected, it means that H_0: $\mu_1 = \mu_2 = \mu_3$ is rejected. This technique has two important disadvantages:

(a) The probability of a type I error increases in proportion to the number of pairwise hypotheses of the form H_0: $\mu_i = \mu_j$, $i < j$.

(b If the number of populations, namely c, is extensive, this technique entails testing $\binom{c}{2}$ pairwise hypotheses. For $c = 4$ there are, for example $\binom{4}{2} = 6$ such hypotheses, while for $c = 8$ there are already $\binom{8}{2} = 28$ such hypotheses. In contrast, the procedure of analysis of variance entails only one statistical test, irrespective of c's value.

In analysis of variance the different populations are usually described as **treatments**. The **observations** on which the analysis is based are then the results that are obtained by applying these treatments to so-called **experimental units**. In the situation where the industrialist wishes to compare the productivity of three groups of workers who have undergone different training programmes, the training programmes can be regarded as the treatments. The worker groups, on the other hand, are the experimental units and each worker's productivity can, for example, be measured in terms of the number of items that are produced per day either by him personally, or under his supervision. The hypothesis being tested, is the equality of the mean number of items being produced daily by the three groups of workers. If these means do not differ, the industrialist reaches the conclusion that the three treatments, or in other words training programmes, do not differ. However, such a conclusion can only be justified if, other than stochastic variations and the effect of training programmes, there are no **assignable sources of variation**. If the workers who underwent one specific training programme work predominantly in a new plant with new equipment, the higher productivity of these workers could be ascribed to higher quality production facilities rather than to a better training programme.

It is therefore important to eliminate assignable sources of variation in the design of the experiment so that any systematic effect in the observations can be ascribed solely to differences between treatments. Even if the treatments do not differ, there will still be differences in the observations as a result of inherent incidental differences between the experimental units. These differences are known as **experimental errors** and play an important role in testing equal treatment means. It is, for example, obvious that there would be differences in worker productivity even if the three training procedures were identical. This can be ascribed to the inherent differences in the abilities of individual workers.

Divergent research situations in which possible differences between treatments have to be tested repeatedly and as accurately as possible, result in different experimental designs. The meaningful design of experiments is, in

fact, a comprehensive subject and its theoretical basis has been extensively documented. The purpose of this chapter is therefore not to provide an overview of this field, but rather to serve as an introduction to the subject. Only a few of the simplest experimental designs that are currently in use in the general field of analysis of variance will be discussed.

Since R.A. Fisher laid the foundation for experimental design in 1935, the theory of experimental design has been expanded considerably. Although analysis of variance was originally designed to solve research problems specifically in the field of agriculture and the biological sciences, it is now used fairly freely in all disciplines.

2 The completely randomized design

2.1 Testing for one or more differences between a number of treatments

Consider once more the situation of the industrialist who wishes to determine whether the productivity of workers who have undergone one of three disparate training programmes differs. Suppose that he employed 24 new workers at the beginning of 1993. In order to compare training programmes, he decides to expose eight new workers to each of these programmes. To eliminate the possibility of any of the training programmes being either prejudiced or favoured, he uses random allocation of the workers to the training programmes. This method is a typical application of a **completely randomized design.**

> A design in which ch experimental units are randomly assigned to c treatments so that each treatment is applied to h units, is known as a **completely randomized design**.

The randomized assignment of the experimental units to the treatments often eliminates systematic effects that are inherent to the experimental units before treatment. If there is little or no information on the nature or composition of the units, this design yields an objective test for

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_c$$

after applying the treatments. The observations of a completely randomized experiment can be represented schematically as in Table 14.1.

2.1.1 Notation

Suppose the total number of observations is $n = ch$. The sum of the n observations is obtained as

$$t = \sum_{i=1}^{c} \sum_{j=1}^{h} x_{ij}$$

Table 14.1 Observations of a completely randomized experiment

Treatments	Observations	Total	Average
1	$x_{11}\ x_{12}\ \ldots\ x_{1h}$	b_1	$\bar{b}_1$
2	$x_{21}\ x_{22}\ \ldots\ x_{2h}$	b_2	$\bar{b}_2$
.	. . . . .	.	.
.	. . . . .	.	.
.	. . . . .	.	.
c	$x_{c1}\ x_{c2}\ \ldots\ x_{ch}$	b_c	$\bar{b}_c$
		t	$\bar{x}$

so that the arithmetic mean of all the observations is given by

$$\bar{x} = \frac{t}{n}$$

If the sum of the h observations (repetitions) of treatment i is now represented by the symbol b_i, it then follows that the corresponding arithmetic mean is given by

$$\bar{b}_i = \frac{b_i}{h},\ i = 1, 2, \ldots, c.$$

With regard to the observations in Table 14.1 it can now be shown that

$$\sum_{i=1}^{c}\sum_{j=1}^{h}(x_{ij} - \bar{x})^2 = h\sum_{i=1}^{c}(\bar{b}_i - \bar{x})^2 + \sum_{i=1}^{c}\sum_{j=1}^{h}(x_{ij} - \bar{b}_i)^2$$

or, briefly, $S_t = S_b + S_e$

The **total sum of squares** S_t can therefore be divided into two components, namely

S_b = **treatment sum of squares** and

S_e = **error sum of squares.**

Large differences between the c treatments will result in large differences in the treatment averages $\bar{b}_1, \bar{b}_2, \ldots, \bar{b}_c$ which will give rise to a large S_b value. Since S_e measures the total mutual variation in the observations of every treatment, the latter is a measure of experimental errors. Furthermore, the expressions

$$\frac{S_b}{c-1} \text{ and } \frac{S_e}{n-c}$$

are known respectively as the **mean treatment sum of squares** and the **mean error sum of squares**.

Assumptions of inference in a completely randomized experiment: Assume that the h observations of the first treatment can be regarded as a random sample of observations from a $n(\mu_1, \sigma^2)$ population. Similarly the h observations of the second treatment are regarded as a random sample of observations from a $n(\mu_2, \sigma^2)$ population, ... and the h observations for the cth treatment as a random sample from a $n(\mu_c, \sigma^2)$ population. We further assume that the observations taken from all the c treatments are independent of each other.

2.1.2 Test procedure

Now consider the testing of

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_c$$

against

H_1: The μ's differ from one another.

If H_0 is true and the inference assumptions for the completely randomized experiment are met,

$$F = \frac{\text{Mean treatment sum of squares}}{\text{Mean error sum of squares}}$$

$$= \frac{S_b/(c-1)}{S_e/(n-c)}$$

has the $F(c-1; n-c)$ distribution.

A large observed value f of F will be an indication of differences in the treatments, which will result in the rejection of H_0. From this the following decision rule follows:

Reject H_0 if $f \geq F_{c-1, n-c; \alpha}$

The test procedure is then set out in Table 14.2 as the so-called **analysis of variance table**. The **degrees of freedom** associated with every sum of squares are obtained as follows:

The total number of degrees of freedom = total number of observations – 1
$= n - 1$

The degrees of freedom of treatments = total number of treatments – 1
$= c - 1$

Since

total number of degrees of freedom = degrees of freedom of treatments + degrees of freedom of errors,

it follows that

degrees of freedom of errors $= (n-1) - (c-1)$
$= n - c$

The analysis of variance table provides a summary of all the calculated values that are required for the calculation of f and is always shown for this reason.

2.1.3 Calculation formulas

The simplest calculation formulas for S_t, S_b and S_e are respectively

$$S_t = \sum_{i=1}^{c} \sum_{j=1}^{h} x_{ij}^2 - \frac{t^2}{n}$$

$$S_b = \frac{1}{h} \sum_{i=1}^{c} b_i^2 - \frac{t^2}{n}$$

$$S_e = S_t - S_b$$

Table 14.2 Analysis of variance table for testing H_0: $\mu_1 = \mu_2 = \ldots = \mu_c$ for a completely randomized experiment

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F
Treatments	S_b	$c - 1$	$S_b/(c-1)$	$\frac{S_b/(c-1)}{S_e/(n-c)}$
Errors	S_e	$n - c$	$S_e/(n-c)$	
Total	S_t	$n - 1$		

EXAMPLE 14.1

A boutique that specializes in exclusive ladieswear allows for three methods of payment by clients, namely cash, credit card or cheque. The manageress would like to determine whether there is any significant difference in the average amounts spent by clients using the respective methods of payment. Table 14.3 shows the amounts (to the nearest rand) for three random samples of size 12, taken from all the transactions over a period of 30 days where settlement has been either in cash, or by credit card or cheque. Indicate the population means of the cash, credit card and cheque sales transactions by μ_1, μ_2 and μ_3 respectively. Although the experimental units, namely the individuals purchasing clothing from the boutique, cannot be 'assigned' to the respective treatments (cash, credit card and cheque) this

experimental design can still be regarded as a completely randomized design, the reason being that the sales transactions in this example have been drawn randomly and were not selected.

Table 14.3 Clothing purchased (in rand) by 36 individuals

Treatments	Observations	Total	Average
1 Cash	230 85 70 170 235 81 97 130 260 190 200 65	1 813	151.08
2 Credit card	385 215 250 350 280 210 230 420 495 260 190 300	3 585	298.75
3 Cheque	185 150 130 240 220 185 170 230 245 193 175 190	2 313	192.75
	$\Sigma\Sigma x_{ij}^2 = 1\,959\,569$	7 711	214.19

From the details contained in Table 14.3 it follows that

$$S_t = \sum_{i=1}^{3}\sum_{j=1}^{12} x_{ij}^2 - \frac{t^2}{n}$$

$$= 230^2 + 85^2 + 70^2 + \ldots + 190^2 - \frac{(7\,711)^2}{36}$$

$$= 1\,959\,569 - 1\,651\,653.4$$

$$= 307\,915.6$$

$$S_b = \frac{1}{h}\sum_{i=1}^{3} b_i^2 - \frac{t^2}{n}$$

$$= \tfrac{1}{12}\{(1\,813)^2 + (3\,585)^2 + (2\,313)^2\} - \frac{(7\,711)^2}{36}$$

$$= 1\,790\,763.6 - 1\,651\,653.4$$

$$= 139\,110.2$$

$$S_e = S_t - S_b$$

$$= 307\,915.6 - 139\,110.2$$

$$= 168\,805.4$$

These results will now be used to compile the analysis of variance table (see Table 14.4) to test

H_0: $\mu_1 = \mu_2 = \mu_3$

against

H_1: The μ's differ.

Table 14.4 Analysis of variance table for testing the equality of the mean amounts spent on clothing and paid for by one of the three methods

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	f
Treatments	139 110.2	2	69 551.1	13.597
Errors	168 805.4	33	5 115.3	
Total	307 915.6	35		

In terms of H_0

$F = \dfrac{S_b/2}{S_e/33}$ has a F(2, 33) distribution.

From Table B.5 $F_{2,\,33;\,0.01} \simeq 5.30$. Since $f > 5.30$, H_0 is rejected at the 1% level of significance. We can therefore conclude that the mean amounts spent and paid for by the three methods differ.

EXAMPLE 14.2

Twenty-eight students register for a particular semester course at a college. In order to compare four different textbooks covering the same topic the students are divided randomly into four classes and subsequently a different textbook is assigned randomly to each of the four classes. Various methods can be used to divide the students into four classes. If twenty-eight identical cards are prepared (seven numbered as 1, seven numbered as 2, seven as 3 and seven as 4), placed in a container and shuffled then each student could draw one. This method leads to a randomized division into classes. Random digits can also be used in various ways to divide the 28 students into the four classes. We shall explain one method: At least 28 successive groups of two digits each are read from Table B.1. Numbers 0 to 24 are now associated with class 1, numbers 25 to 49 with class 2, numbers 50 to 74 with class 3 and numbers 75 to 99 with class 4. If the quota, namely 7, of a particular class is already filled, further random numbers that are associated with that class are ignored. Table 14.5 shows a possible division of the students by using the random digits in Table B.1 (fourth page, row 41, column 1, horizontal). The student number indicates the sequence of registration. Note that the last three random numbers were redundant, since the students had to be allocated to class 2 in any case, to fill the quota of seven.

Table 14.6 now shows the final achievement of the 23 students who wrote the examination in the course. Two students discontinued their studies during the semester, while three others were forbidden to write the examination because of poor class attendance.

Since there is no longer an equal number of observations per treatment, we use the following formulas to calculate S_t, S_b and S_e. Suppose there are h_1 observa-

Table 14.5 Randomized classification of 28 students into four classes, each with 7 students

Student number	Random number	Class	Student number	Random number	Class
1	13	1	15	82	4
2	92	4	16	55	3
3	62	3	17	10	1
4	98	4	18	16	1
5	41	2	19	91	4
6	56	3	20	31	2
7	65	3	21	71	3
8	74	3	22	92	4
9	39	2	23	03	1
10	46	2	24	10	1
11	93	4	25	01	1
12	55	3	26	35	2
13	86	4	27	31	2
14	17	1	28	35	2

Table 14.6 Percentages scored by 23 students in an examination

Treatments	Observations	Total	Average
Text book 1	72 65 81 83 67 73	441	73.50
Text book 2	48 63 71 55 52 49 54	392	56.00
Text book 3	72 76 82 73 69	372	74.40
Text book 4	65 68 75 61 92	361	72.20
	$\Sigma\Sigma x_{ij}^2 = 109\,470$	1 566	68.09

tions of treatment 1, h_2 observations of treatment 2, ... and h_c observations of treatment c, so that $n = \sum_{i=1}^{c} h_i$, it then follows that:

$$S_t = \sum_{i=1}^{c} \sum_{j=1}^{h_i} x_{ij}^2 - \frac{t^2}{n}$$

$$S_b = \sum_{i=1}^{c} \left(\frac{b_i^2}{h_i} \right) - \frac{t^2}{n}$$

where b_i indicates the sum of the h_i observations of treatment i. As demonstrated above it also follows that

$$S_e = S_t - S_b.$$

The sums of the squares S_t, S_b and S_e still have $(n-1)$, $(c-1)$ and $(n-c)$ degrees of freedom respectively. From the data in Table 14.6 it then follows that

$$\begin{aligned} S_t &= (72)^2 + (65)^2 + \ldots + (92)^2 - \frac{(1\,566)^2}{23} \\ &= 109\,470 - 106\,624.17 \\ &= 2\,845.83 \end{aligned}$$

$$S_b = \frac{(441)^2}{6} + \frac{(392)^2}{7} + \frac{(372)^2}{5} + \frac{(361)^2}{5} - \frac{(1\,566)^2}{23}$$

$$= 108\,106.50 - 106\,624.17$$

$$= 1\,482.33$$

$$S_e = 2\,845.83 - 1\,482.33$$

$$= 1\,363.50$$

These results will then be used to compile the analysis of variance table (see Table 14.7) to test

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

against

H_1: the μ's differ.

If H_0 is true

$$F = \frac{S_b/3}{S_e/19}$$

has an F(3, 19) distribution.

Table 14.7 Analysis of variance table for testing the equality of the mean achievements of four classes

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	f
Treatments	1 482.33	3	494.11	6.89
Errors	1 363.50	19	71.76	
Total	2 845.83	22		

From Table B.5 $F_{3,\,19;\,0.01} = 5.01$. Since $f \geq 5.01$, H_0 is rejected at the 1% level of significance and we can conclude that the four textbooks do, in fact, differ.

2.2 Scheffé's method of pairwise comparisons

If the hypothesis

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_c$$

is rejected, it is often important for the researcher to determine which specific population means differ from one another. As stated in Section 1, successive testing of the hypotheses of the form $H_0: \mu_i = \mu_j$, $i < j$ will probably lead to a large type I error.

Techniques have been developed to overcome this problem. Using these, all possible comparisons of treatment averages can be done pairwise without

exceeding the predetermined level of significance. We shall confine this discussion to the Scheffé procedure.

Scheffé's procedure of **paired comparisons** of the treatment averages $\bar{b}_1, \bar{b}_2, \ldots, \bar{b}_c$ in a completely randomized experiment is based on the following set of confidence intervals for $\mu_i - \mu_j$, $i < j$:

$$\left\{(\bar{b}_i - \bar{b}_j) - \sqrt{m_{ij}},\ (\bar{b}_i - \bar{b}_j) + \sqrt{m_{ij}}\right\},$$

in which $m_{ij} = \left(\frac{1}{h_i} + \frac{1}{h_j}\right)\left(\frac{c-1}{n-c}\right) S_e F_{c-1,\, n-c;\, 0.05}$

The joint confidence coefficient for these intervals is 0.95 and we can conclude that $\bar{b}_i$ and $\bar{b}_j$ differ significantly if the calculated interval for $\mu_i - \mu_j$ excludes the number 0.

Note that in the case of an equal number of observations per treatment, in other words if $h_1 = h_2 = \ldots = h_c = h$, then $m_{ij} = m$ for all i and j, with

$$m = \frac{2(c-1)}{h(n-c)} S_e F_{c-1,\, n-c;\, 0.05}$$

EXAMPLE 14.3

In Example 14.2 we reached the conclusion that the class means differ for students using different textbooks. Pairwise comparisons of $\bar{b}_1$, $\bar{b}_2$, $\bar{b}_3$ and $\bar{b}_4$ entail the calculation of all possible confidence intervals for $\mu_i - \mu_j$, $i < j$.

First consider the calculation of a confidence interval for $\mu_1 - \mu_2$, namely

$$\left\{(\bar{b}_1 - \bar{b}_2) - \sqrt{m_{12}},\ (\bar{b}_1 - \bar{b}_2) + \sqrt{m_{12}}\right\}$$

From Table 14.6 it follows that $\bar{b}_1 = 73.50$, $\bar{b}_2 = 56.00$, $h_1 = 6$ and $h_2 = 7$. Since $S_e = 1\,363.50$ (see Table 14.7) and $F_{3,\,19;\,0.05} = 3.13$ it follows that

$$\begin{aligned} m_{12} &= \left(\frac{1}{6} + \frac{1}{7}\right)\left\{\left(\frac{3}{19}\right)(1\,363.50)(3.13)\right\} \\ &= 0.3095 \times 673.856 \\ &= 208.56 \end{aligned}$$

and the required confidence interval is

$$\begin{aligned} &\left\{(73.50 - 56.00) - \sqrt{208.56},\ (73.50 - 56.00) + \sqrt{208.56}\right\} \\ &= (3.06,\ 31.94) \end{aligned}$$

Similarly, confidence intervals can be calculated for all the other differences in treatment means. Table 14.8 gives a summary of the calculated intervals. If an interval for $(\mu_i - \mu_j)$ includes zero, this means that $(\mu_i - \mu_j)$ does not differ

significantly from zero. In other words H_0: $\mu_i = \mu_j$ is not rejected (see Remark (b), Chapter 12, Section 4). The joint probability that at least one of the pairwise equality hypotheses

$$H_0: \mu_1 = \mu_2;\ H_0: \mu_1 = \mu_3;\ H_0: \mu_1 = \mu_4$$
$$H_0: \mu_2 = \mu_3;\ H_0: \mu_2 = \mu_4;\ H_0: \mu_3 = \mu_4$$

is incorrectly indicated as invalid, is therefore 0.05.

Table 14.8 95% confidence intervals according to Scheffé's method for pairwise comparison of the treatment averages given in Example 14.2

$\mu_i - \mu_j$	$\sqrt{m_{ij}}$	Interval for $(\mu_i - \mu_j)$	Decision
$\mu_1 - \mu_2$	14.44	(3.06, 31.94)	μ_1 and μ_2 differ
$\mu_1 - \mu_3$	15.72	(–16.62, 14.82)	μ_1 and μ_3 do not differ
$\mu_1 - \mu_4$	15.72	(–14.42, 17.02)	μ_1 and μ_4 do not differ
$\mu_2 - \mu_3$	15.20	(–33.60, –3.20)	μ_2 and μ_3 differ
$\mu_2 - \mu_4$	15.20	(–31.40, –1.00)	μ_2 and μ_4 differ
$\mu_3 - \mu_4$	16.42	(–14.22, 18.62)	μ_3 and μ_4 do not differ

It follows from Table 14.8 that Textbook 2 differs from the other three textbooks. The poor performance of the group for whom this book was prescribed, cannot be dismissed as incidental.

Notes

(a) **Tukey** suggested a pairwise comparison technique that is similar to Scheffé's. Although Tukey's procedure is based on confidence intervals that are generally shorter than the corresponding Scheffé intervals and therefore indicate significant differences between means more readily, it has the serious shortcoming that it can be applied only if there are an equal number of observations per treatment. The SAS program series offers the possibility of using other multiple comparison techniques in addition to those of Tukey and Scheffé.

(b) Pairwise comparisons form a special case of the **multiple comparison technique** in terms of which confidence intervals with a prescribed confidence coefficient can be determined for all possible ***contrasts*** $\sum_{i=1}^{c} k_i\mu_i$ with $k_1, k_2, \ldots, k_c$ chosen so that $\sum_{i=1}^{c} k_i = 0$. Through the correct choice of $k_1, k_2, \ldots, k_c$ it is, for example, possible to decide whether the means of two or more groups of treatment differ from one another.

(c) In Section 6.3 of Chapter 12 a procedure was discussed for determining whether or not there are differences between two population variances.

Bartlett's test can be used to test the assumption of equal variances of the treatment populations where more than two treatments are being considered. It is important to bear in mind that both these tests are based on the assumption that the c samples are independent of one another and that they originate from normal populations.

(d) Tests for normality can be used to investigate the assumption of normality of the c treatment populations (see Section 2, Chapter 15). If each observation is standardized by

$$z_{ij} = \frac{x_{ij} - \bar{b}_i}{s_i}; i = 1, 2, \ldots, c; j = 1, 2, \ldots, h_i$$

in which s_i is the standard deviation of $x_{i1}, x_{i2}, \ldots, x_{ih_i}$, the test for normality can be performed on the joint sample of $\sum_{i=1}^{c} h_i$ z values.

3 The randomized block design

The completely randomized design can be regarded as the simplest and most basic experimental design. However, it is often found that the observations on experimental units that are allocated to one or more specific treatments in terms of this design, differ substantially from the other observations even when there are no real differences between the treatments. This is brought about by the presence of one or more additional sources of variation. As a result of these sources or external influences, the experimental units are not **homogeneous** even before treatment, but differ from one another.

In the situation outlined in Example 14.2, most of the randomly assigned students in group 2 could well be students whose academic performance has always been poor or average. In such a case, the relatively poor performance of this group in comparison with the other groups would not necessarily be an indication of a poor quality textbook. Although the random assignment of students to class groups will usually provide a mix of good, average and poor students in each class group, it is safer to divide the students into homogeneous **blocks** in terms of academic performance beforehand. Since four class groups need to be formed, we also choose block sizes of four, so that each block of four students represents a group of students of equal ability. The four students in each block are then assigned randomly to the four different class groups. This design is known as a randomized block design.

A design in which ch heterogeneous experimental units are divided into h homogeneous blocks of size c and the units in each block are then assigned randomly to the c different treatments, is known as a **randomized block design**.

The observations can now be represented schematically as in Table 14.9.

Table 14.9 Observations of a randomized block experiment

Treatments	Blocks 1 2 ... h	Total	Average
1	$x_{11}\ x_{12}\ \cdots\ x_{1h}$	b_1	$\bar{b}_1$
2	$x_{21}\ x_{22}\ \cdots\ x_{2h}$	b_2	$\bar{b}_2$
.	. . ... .	.	.
.	. . ... .	.	.
.	. . ... .	.	.
c	$x_{c1}\ x_{c2}\ \cdots\ x_{ch}$	b_c	$\bar{b}_c$
Total	$k_1\ k_2\ \cdots\ k_h$	t	
Average	$\bar{k}_1\ \bar{k}_2\ \cdots\ \bar{k}_h$		$\bar{x}$

3.1 Notation

As in the previous case, t denotes the sum of the $n = ch$ observations and $\bar{x} = t/n$ is the arithmetic mean of all the n observations. The block totals (column totals) and block averages (column averages) are given by $k_1, k_2, \ldots, k_h$ and $\bar{k}_1, \bar{k}_2, \ldots, \bar{k}_h$ where $\bar{k}_i = k_i/c$. Similarly, it follows that the treatment totals and treatment averages are given respectively by $b_1, b_2, \ldots, b_c$ and $\bar{b}_1, \bar{b}_2, \ldots, \bar{b}_c$, in which $\bar{b}_i = b_i/h$.

It can now be shown that the total sum of squares can be divided into three components, namely

$$\sum_{i=1}^{c}\sum_{j=1}^{h}(x_{ij} - \bar{x})^2 = h\sum_{i=1}^{c}(\bar{b}_i - \bar{x})^2 + c\sum_{j=1}^{h}(\bar{k}_j - \bar{x})^2 + \sum_{i=1}^{c}\sum_{j=1}^{h}(x_{ij} - \bar{b}_i - \bar{k}_j + \bar{x})^2$$

or, briefly, $S_t = S_b + S_k + S_e$.

The **total sum of squares** S_t now consists of three components, namely

S_b = **treatment sum of squares**

S_k = **block sum of squares**

S_e = **error sum of squares**.

Note that the definition of S_e now differs from its definition in the case of the completely randomized design. We can also write S_e as

$$S_e = \sum_{i=1}^{c}\sum_{j=1}^{h}\{(x_{ij} - \bar{b}_i) - (\bar{k}_j - \bar{x})\}^2$$

What actually happens, is that the effect of the block differences is eliminated from the error sum of squares, so that S_e is only a measure of experimental errors.

Suppose that the observations can be regarded as observed values of n independent normal variables with equal variances. In order to test

H_0: the treatment effects are the same

against

H_1: the treatment effects differ,

the fact used is that

$$F = \frac{\text{Mean treatment sum of squares}}{\text{Mean error sum of squares}}$$

$$= \frac{S_b/(c-1)}{S_e/(c-1)(h-1)}$$

has the $F\{c-1, (c-1)(h-1)\}$ distribution.

Note that it is also possible to test

H_0: The block effects are the same

against

H_1: The block effects differ.

In this case it must be noted that

$$F = \frac{\text{Mean block sum of squares}}{\text{Mean error sum of squares}}$$

$$= \frac{S_k/(h-1)}{S_e/(c-1)(h-1)}$$

has the $F\{h-1, (c-1)(h-1)\}$ distribution.

The analysis of variance table is given in Table 14.10.

3.2 Calculation formulas

The simplest calculation formulas for S_t, S_b, S_k and S_e are respectively

$$S_t = \sum_{i=1}^{c}\sum_{j=1}^{h} x_{ij}^2 - \frac{t^2}{n}$$

$$S_b = \frac{1}{h}\sum_{i=1}^{c} b_i^2 - \frac{t^2}{n}$$

$$S_k = \frac{1}{c}\sum_{j=1}^{h} k_j^2 - \frac{t^2}{n}$$

$$S_e = S_t - S_b - S_k$$

As in the completely randomized design, H_0 is rejected if f is greater than or equal to the critical value.

Table 14.10 Analysis of variance table for the analysis of a randomized block experiment

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F
Treatments	S_b	$c - 1$	$S_b/(c-1)$	$\dfrac{S_b/(c-1)}{S_e/(c-1)(h-1)}$
Blocks	S_k	$h - 1$	$S_k/(h-1)$	$\dfrac{S_k/(h-1)}{S_e/(c-1)(h-1)}$
Errors	S_e	$(c-1)(h-1)$	$S_e/(c-1)(h-1)$	
Total	S_t	$n - 1$		

EXAMPLE 14.4

Suppose we wish to compare the average valuation of houses in the three areas of the Witwatersrand, namely the Johannesburg area, the East Rand and the West Rand. In order to ensure that differences in valuations are not due to differences in house size, we shall use a randomized block design. Three houses of comparable size are repeatedly selected randomly from the three areas and the valuations are noted. Each group of three houses forms a block. Table 14.11 reflects the valuations of 33 houses on the Witwatersrand that have been obtained in this way. The total floor area is shown in brackets below the valuations, from which it can be seen that the block units are homogeneous.

From the data in Table 14.11 it follows that

$$S_t = \sum_{i=1}^{3}\sum_{j=1}^{11} x_{ij}^2 - \frac{t^2}{33}$$

$$= 1\,849\,042 - \frac{(7\,230)^2}{33}$$

$$= 265\,014.7$$

Table 14.11 Valuations (R1 000) for 33 houses on the Witwatersrand. The total floor area (m^2) is indicated in brackets

	Blocks						
Treatments	1	2	3	4	5	6	7
1 Johannesburg area	320 (387)	160 (160)	195 (163)	165 (134)	300 (183)	315 (190)	120 (166)
2 East Rand	430 (390)	105 (160)	170 (162)	160 (134)	206 (183)	190 (189)	165 (167)
3 West Rand	515 (376)	185 (160)	95 (184)	185 (134)	220 (185)	160 (188)	205 (168)
Total, k_i	1 265	450	460	510	726	665	490
Average, $\bar{k}_i$	421.7	150.0	153.3	170.0	242.0	221.7	163.3

Table 14.11 (continued)

Treatments	Blocks				Total b_i	Average $\bar{b}_i$
	8	9	10	11		
1 Johannesburg area	170 (212)	120 (175)	300 (183)	210 (214)	2 375	215.9
2 East Rand	215 (212)	159 (174)	150 (182)	320 (213)	2 270	206.4
3 West Rand	240 (210)	305 (174)	240 (180)	235 (213)	2 585	235.0
Total, k_j	625	584	690	765	t = 7 230	
Average, $\bar{k}_j$	208.3	194.7	230.0	255.0		$\bar{x}$ = 219.1

$c = 3$; $h = 11$; $n = ch = 33$; $\Sigma\Sigma x_{ij}^2 = 1\,849\,042$

$$S_b = \frac{1}{11}\sum_{i=1}^{3} b_i^2 - \frac{t^2}{33}$$

$$= \frac{1}{11}\{(2\,375)^2 + (2\,270)^2 + (2\,585)^2\} - \frac{(7\,230)^2}{33}$$

$$= 1\,588\,704.5 - 1\,584\,027.3$$

$$= 4\,677.2$$

$$S_k = \frac{1}{3}\sum_{j=1}^{11} k_j^2 - \frac{t^2}{33}$$

$$= \frac{1}{3}\{(1\,265)^2 + (450)^2 + \ldots + (765)^2\} - \frac{(7\,230)^2}{33}$$

$$= 1\,758\,944 - 1\,584\,027.3$$

$$= 174\,916.7$$

$$S_e = S_t - S_b - S_k$$

$$= 265\,014.7 - 4\,677.2 - 174\,916.7$$

$$= 85\,420.8$$

These results will now be used to compile the analysis of variance table (see Table 14.12) for testing

(a) H_0: The treatment effects are the same
against
H_1: The treatment effects differ

(b) H_0: The block effects are the same
against
H_1: The block effects differ.

Table 14.12 Analysis of variance table for testing the equality of the mean building valuations for the three Witwatersrand areas

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	f
Treatments	4 677.2	2	2 338.6	0.55
Blocks	174 916.7	10	17 491.7	4.10
Errors	85 420.8	20	4 271.0	
Total	265 014.7	32		

If H_0 as given in (a) is valid,

$$F = \frac{S_b/2}{S_e/20}$$

has the F(2, 20) distribution. From Table B.5, $F_{2,\,20;\,0.05} = 3.49$ and since $f < 3.49$, it follows that H_0 is not rejected at a 5% level of significance. We can therefore conclude that the mean building valuations of houses of comparable floor area in the Johannesburg area, the East Rand and the West Rand do not differ from one another.

If H_0 applies as in case (b),

$$F = \frac{S_k/10}{S_e/20}$$

has the F(10, 20) distribution. From Table B.5, $F_{10,\,20;\,0.05} = 2.35$ and since $f \geq 2.35$ H_0 is rejected, so that the eleven block averages actually differ significantly from one another at a 5% level of significance (in fact, also at a 1% level of significance). This indicates that the inclusion of blocks was in fact justified and that this led to a significant decrease in the mean error sum of squares.

Notes

(a) A randomized block design can be regarded as an extension of the t test for paired observations to the case of more than two treatments. Example 14.4 gives a good illustration of this for $c = 3$. The randomized block design can also be regarded as the problem of c associated samples (also see Chapter 16, Section 4).

(b) Scheffé's procedure of pairwise comparison of $\bar{b}_1, \bar{b}_2, \ldots, \bar{b}_c$, is also valid for the randomized block design. In this case $m_{ij} = m$ for all i and j namely

$$m = \frac{2S_e}{h(h-1)} F_{c-1,\,(c-1)(h-1);\,0.05}$$

with $S_e = S_t - S_b - S_k$

(c) The above randomized block design required only one observation per (block) × (treatment) combination. This design can be extended to the case of m observations per (block) × (treatment) combination.

(d) If the heterogeneity of the experimental units is caused by (or related to) a continuous measurable stochastic variable Y, **analysis of covariance** can be used as an alternative for the randomized block design. In terms of this method y is included as a **co-factor** in the analysis and each x observation is corrected beforehand for the influence of the co-factor before testing for differences in treatment means.

4 Factorial designs

In the designs that we have discussed thus far in this chapter, we have repeatedly compared alternative treatments with one another. In Example 14.1 the three 'treatments' were the methods of payment, whereas the four different textbooks were regarded as the so-called treatments in Example 14.2. Different treatments can also be regarded as the **levels of a factor.** Thus the method of payment that was considered in Example 14.1 is the factor with cash, credit cards and cheques as the levels. In Example 14.2 'textbook' was the factor, and the four specific books that were compared form the four levels. Both the completely randomized design and the randomized block design are one-factor designs, since the data collected is used only to test whether the levels of a single factor differ from one another. If the levels of a factor do differ significantly from one another, such a factor is considered to be significant.

Often, more than one factor in a research project is important and time and costs can be saved if different factors can be tested for significance within a single experiment.

EXAMPLE 14.5

A company that markets milk powder in cooperation with a hypermarket plans to conduct an investigation into the popularity of five competitive **brands of milk powder**. They also want to determine simultaneously whether certain **packaging sizes** are more popular than others. In order to realize both objectives, they decide to keep stocks of all five brands of milk powder in packagings of 5 kg, 2 kg, 1 kg and 500 g on the shelves of this hypermarket. On ten randomly chosen weekdays during a period of one month the monetary value of the sales of each of the twenty (milk powder brand) × (packaging size) combinations was observed. This design is a two-factor design and is known as a (5 × 4) factorial design with ten observations per combination of the 2 factors. Each combination of milk powder and packaging size is known as a **cell**.

EXAMPLE 14.6

An agricultural research organization plans to compare three **maize varieties** with one another. To save costs, the researchers decide to investigate the effect of four

fertilizers and two **planting methods**. Since a large number of fairly homogeneous agricultural plots with a size of $\frac{1}{2}$ ha each are available, the researchers decide on random allocation of five plots for each (variety) × (fertilizer) × (planting method) combination. These plots are then treated beforehand with the respective fertilizers, after which the different varieties are planted in the appropriate manner. After a certain period the yield of each plot is determined. This is a factorial experiment with three factors namely, variety, fertilizer type and planting method. Briefly this can be classified as a $(3 \times 4 \times 2)$ factorial design with five observations for each combination of factors.

In this section only the design and analysis of two-factor experiments will be discussed. The extension to three-factor or multifactor experiments is fairly direct.

Suppose factor A has r levels and factor B has k levels. The design in which rkh experimental units are assigned randomly to the rk cells so that every combination of treatments is applied to h units, is known as an $(r \times k)$ **factorial design with h repetitions per cell**.

There are various important advantages attached to the use of an $(r \times k)$ factorial design, namely:

(a) It is far more advantageous than the use of, for example, two one-factor completely randomized designs because of savings in terms of time and costs.

(b) Each factor is not merely analyzed at one level of the other factor as is the case in a completely randomized design. The analysis is done under a variety of circumstances (the different levels of the other factor). This makes the conclusions derived from the analysis of each factor more generally acceptable.

(c) However, the most important benefit attached to a factorial experiment is that this makes the study of **interaction** or interplay between different factors possible.

To elucidate advantage (c) even further, we look at the following examples: With regard to Example 14.5 one or more (milk powder) × (packaging) combinations could well be exceptionally popular, while very few people buy other factorial combinations. Moreover, in the situation set out in Example 14.6, a specific (fertilizer) × (planting method) combination could result in one maize variety yielding hardly any crop. However, testing for interaction effects is possible only in factorial experiments in which more than one factor is studied simultaneously and where the number of repetitions per cell is larger than one.

The observations of an $(r \times k)$ factorial experiment can be represented as in Table 14.13:

Table 14.13 Observations of an $(r \times k)$ factorial experiment

		Factor B				Total	Average
		1	2	...	k		
Factor A	1	h x values, sum = s_{11}	h x values, sum = s_{12}	...	h x values, sum = s_{1k}	a_1	$\bar{a}_1$
	2	h x values, sum = s_{21}	h x values, sum = s_{22}	...	h x values, sum = s_{2k}	a_2	$\bar{a}_2$
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	r	h x values, sum = s_{r1}	h x values, sum = s_{r2}	...	h x values, sum = s_{rk}	a_r	$\bar{a}_r$
Total		b_1	b_2	...	b_k	t	
Average		$\bar{b}_1$	$\bar{b}_2$	...	$\bar{b}_k$		$\bar{x}$

4.1 Notation

Suppose $n = rkh$ is the total number of observations (x values) and that t is the sum of these n observations, then it follows that the arithmetic mean of all the observations will be given by

$$\bar{x} = \frac{t}{n}$$

Suppose further that s_{ij} is the sum of the h observations in cell (i, j) and indicate the average of this cell by $\bar{s}_{ij}$ then it follows that

$$\bar{s}_{ij} = \frac{s_{ij}}{h}$$

Now indicate the sum of the kh observations at level i of factor A by a_i. The corresponding averages of the r levels of A are then

$$\bar{a}_i = \frac{a_i}{kh}, i = 1, 2, ..., r$$

Similarly it follows for b_j, the sum of the rh observations at level j of factor B, that

$$\bar{b}_j = \frac{b_j}{rh}, j = 1, 2, ..., k$$

If $h > 1$ the total sum of squares can be divided into four components as follows:

$$S_t = S_a + S_b + S_{ab} + S_e$$

where

S_t = **total sum of squares,**

S_a = **factor A sum of squares,**

S_b = **factor B sum of squares,**

S_{ab} = **interaction sum of squares** and

S_e = **error sum of squares.**

The easiest method for calculating the respective sums of squares, is as follows:

$$S_t = \sum_{\text{all } x} x^2 - \frac{t^2}{n}$$

$$S_a = \frac{1}{kh} \sum_{i=1}^{r} a_i^2 - \frac{t^2}{n}$$

$$S_b = \frac{1}{rh} \sum_{j=1}^{k} b_j^2 - \frac{t^2}{n}$$

$$S_{ab} = \frac{1}{h} \sum_{i=1}^{r} \sum_{j=1}^{k} s_{ij}^2 - \frac{t^2}{n} - S_a - S_b$$

$$S_e = S_t - S_a - S_b - S_{ab}$$

Large values for S_a and S_b respectively indicate possible differences between the mean levels of factor A and between the mean levels of factor B, whereas a large value for S_{ab} indicates an interaction between the two factors.

If the observations can be regarded as observed values of n independent normal variables with equal variances, the following hypotheses can be tested:

$H_0(a)$: The effects of the levels of factor A are the same

against

$H_1(a)$: The effects of the levels of factor A differ.

$H_0(b)$: The effects of the levels of factor B are the same

against

$H_1(b)$: The effects of the levels of factor B differ.

$H_0(ab)$: There is no interaction between factors A and B

against

$H_1(ab)$: There is an interaction between factor A and factor B.

A summary of all the statistics required for a factorial experiment is given in Table 14.14. On the basis of the completely randomized design and randomized block design the reader should be able to determine the degrees of freedom for each F test on the assumption that the hypotheses that have been tested, are in fact valid.

Table 14.14 Analysis of variance table for analysis of the $(r \times k)$ factorial experiment $(h > 1)$

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F
Factor A	S_a	$r - 1$	$S_a/(r - 1)$	$\dfrac{S_a/(r-1)}{S_e/(n-rk)}$
Factor B	S_b	$k - 1$	$S_b/(k - 1)$	$\dfrac{S_b/(k-1)}{S_e/(n-rk)}$
Interaction AB	S_{ab}	$(r - 1)(k - 1)$	$S_{ab}/(r - 1)(k - 1)$	$\dfrac{S_{ab}/(r-1)(k-1)}{S_e/(n-rk)}$
Error	S_e	$n - rk$	$S_e/(n - rk)$	
Total	S_t	$n - 1$		

Notes

In the case where only one observation per cell was taken, in other words where $h = 1, S_e = 0$. Here it is not possible to test $H_0(ab)$. In such a case S_{ab} is regarded as the error sum of squares and the F values for testing $H_0(a)$ and $H_0(b)$ are obtained by replacing $S_e/(n-rk)$ with $S_{ab}/(r-1)(k-1)$. The denominator's degrees of freedom for $h = 1$ are then $(r-1)(k-1)$.

EXAMPLE 14.7

Two of the factors that might affect the number of fungi counted on the leaves of an avocado pear tree are:

- Factor A: Concentration of the fungicide used to spray the leaves, namely, a 10% or 30% solution of copper sulphate.
- Factor B: The position (north, south, east or west) of the section of the tree from which leaves are selected for examination.

Table 14.15 reflects the data obtained after random choice of four trees for each of the eight treatment combinations and fungus counts.

Table 14.15 Number of fungi counted on the leaves of 32 avocado pear trees. Cell totals are shown in brackets

Fungicide (Factor A)	Position (Factor B)				Total	Average
	North	South	East	West	a_i	$\bar{a}_i$
10% solution	93 78 85 81(337)	69 64 61 58(252)	84 65 79 81(309)	73 76 71 77(297)	1 195	74.69
30% solution	103 108 107 111(429)	49 44 38 41(172)	67 73 84 82(306)	80 81 85 78(324)	1 231	76.94
Total, b_i **Average, $\bar{b}_i$**	766 95.75	424 53.00	615 76.88	621 77.63	t = 2 426	$\bar{x}$ = 75.81

$r = 2; k = 4; h = 4; n = rkh = 32; \sum\sum s_{ij}^2 = 773\,000; \sum_{\text{all } x} x^2 = 193\,992$

$$S_t = \sum_{\text{all } x} x^2 - \frac{t^2}{n}$$

$$= (93)^2 + (69)^2 + \ldots + (78)^2 - \frac{(2\,426)^2}{32}$$

$$= 193\,992 - 183\,921.13$$

$$= 10\,070.87$$

$$S_a = \frac{1}{16}\sum_{i=1}^{2} a_i^2 - \frac{t^2}{32}$$

$$= \frac{1}{16}\{(1\,195)^2 + (1\,231)^2\} - \frac{(2\,426)^2}{32}$$

$$= 183\,961.63 - 183\,921.13$$

$$= 40.50$$

$$S_b = \frac{1}{8}\sum_{j=1}^{4} b_j^2 - \frac{t^2}{32}$$

$$= \frac{1}{8}\{(766)^2 + (424)^2 + (615)^2 + (621)^2\} - \frac{(2\,426)^2}{32}$$

$$= 191\,299.75 - 183\,921.13$$

$$= 7\,378.62$$

$$S_{ab} = \frac{1}{4}\sum\sum s_{ij}^2 - \frac{t^2}{32} - S_a - S_b$$

$$= \frac{1}{4}\{(337)^2 + (252)^2 + \ldots + (324)^2\} - \frac{(2\,426)^2}{32} - 40.47 - 7\,378.6$$

$$= 193\,250 - 183\,921.13 - 40.50 - 7\,378.62$$

$$= 1\,909.75$$

$$S_e = S_t - S_a - S_b - S_{ab}$$
$$= 10\,070.87 - 40.50 - 7\,378.62 - 1\,909.75$$
$$= 742$$

Table 14.16 Analysis of variance table for analysing the number of fungi counted on the leaves of 32 avocado pear trees

Source of variance	Sum of squares	Degrees of freedom	Mean sum of squares	f
Concentration (A)	40.50	1	40.50	1.31
Position (B)	7 378.62	3	2 459.54	79.55
Interaction (AB)	1 909.75	3	636.58	20.59
Error	742.00	24	30.92	
Total	10 070.87	31		

From Table B.5 it follows that $F_{1,\,24;\,0.05} = 4.26$ and that $F_{3,\,24;\,0.05} = 3.01$. Since $F = 1.31 < 4.26$, $H_0(a)$ cannot be rejected at the 5% level of significance. The values of F in testing $H_0(b)$ and $H_0(ab)$ lead to the conclusion that both position and interaction affect the findings.

Position	Concentration 10%	Concentration 30%
North	84.25	107.25
South	63.00	43.00
East	77.25	76.50
West	74.25	81.00

From this data one can conclude that trees on the northern side that were treated with a 10% fungicide solution had lower fungus counts that the ones that were treated with a 30% solution. We arrive at precisely the opposite conclusion on comparing the averages on the southern side. This phenomenon explains the significance of the effect of interaction.

Notes

(a) Scheffé's procedure of pairwise comparisons can be used in this case to compare the level averages of factor A and the level averages of factor B with each other. Since all the level averages of factor A are based on the same number of observations, namely kh, it follows that the value of the constant m for the **pairwise comparison of the level averages of factor A**, as defined in Section 2, is given by

$$m = \frac{2(r-1)S_e}{k^2rh(h-1)} F_{r-1,\,n-rk;\,0.05}$$

with $S_e = S_t - S_a - S_b - S_{ab}$.

Similarly, it follows that the constant m for the **pairwise comparison equation of the level averages of factor** B is given by

$$m = \frac{2(k-1)S_e}{r^2kh(h-1)} F_{k-1,\,n-rk;\,0.05}$$

(b) If there is interaction between factors A and B (in other words $H_0(ab)$ is rejected) and $H_0(a)$ is also rejected, then one can conclude that the levels of A differ significantly, but these differences should only be interpreted per level of factor B. The same applies if $H_0(b)$ or both $H_0(a)$ and H_0(b) are rejected in the presence of interaction between A and B.

(c) Although this possibility is not discussed here, pairwise comparison of the cell averages can also be executed by means of Scheffé's procedure.

(d) The $(r \times k)$ factorial design discussed in this section, is known more specifically as the **completely randomized $(r \times k)$ factorial design** with h repetitions per factor combination. By dividing the $n = rkh$ experimental units in h homogeneous blocks of size rk each and assigning each unit in a block randomly to a factor combination, we obtain the so-called **$(r \times k)$ factorial block design** with h blocks.

5 Concluding remarks

5.1 Models for analysis of variance

The sole purpose of analysis of variance is to test the null hypothesis, namely that the effects of the levels of a factor are the same. In factorial experiments the null hypothesis that there is no interaction between different factors, can also be tested. The validity of the F tests that are used to test these hypotheses rests on assumptions regarding the means, variances and underlying distribution of the observations. The theoretical formulation of these assumptions in terms of the unknown population parameters are known as the **model** for a specific design. Although this is not discussed in this chapter as such, each design has a specific model.

5.2 Stochastic effect models

All the designs discussed in this chapter are associated with **fixed effect models**. These models assume that the levels of a factor under consideration represent the complete collection of levels that are of interest to the researcher. However, the levels of a factor should preferably be regarded as a random

sample from a population of treatments or levels. In such a case the effect of each level should preferably be regarded as a stochastic variable. In order to analyse the observations in such an experiment, additional assumptions need to be made. Models that are formulated for these situations, are known as **stochastic effect models**. Suppose, for example, that milk powder is marketed under ten different **trademarks** and in twelve different **packaging sizes**. If five trademarks and four packaging sizes are chosen randomly in an investigation to determine whether the sales of different brands of milk powder and different packaging sizes differ, a (5×4) stochastic effect model is obtained in contrast with the (5×4) fixed effect model in Example 14.5. However, the analysis of observations in cases where stochastic effect models are assumed do not differ much from the analyses considered in this chapter.

5.3 Missing data

On occasion some of the collected data in an experiment could be lost. Thus crop yields from one or more agricultural plots could be detrimentally affected by a locust plague or a hail storm, and the resultant observations would consequently be unusable. On the other hand, workers used in a productivity study might resign or fall ill. In a randomized block or a factorial design such lost observations can lead to the failure of an entire experiment. If only one or a few observations are missing, they can be estimated from the others. Methods to this effect are discussed in various textbooks. It is important to remember that the total number of degrees of freedom must be reduced by the number of missing values that were estimated.

5.4 Validity of assumptions

Researchers should note that the F tests that are used in analysis of variance can lead to inaccurate conclusions if the observations do not meet the required assumptions. It is not always possible to determine the size of the type I error in cases where there are large deviations from these assumptions. One usually assumes that, in the case of small and moderate deviations, this probability will not differ too much from the level of significance. In Notes (c) and (d) of Section 2 the confirmation of these assumptions was briefly considered. However, investigating all the assumptions repeatedly before performing analysis of variance is a tedious process. It is therefore not surprising that, in many cases, the researcher does not subscribe to this form of insurance. Chapter 16 includes discussions on distribution-free tests that can be used for completely randomized designs and randomized block designs in which the assumption of normality is not valid.

5.5 Computer programs

Because of the extent of the calculations in analysis of variance, this type of analysis is seldom done by hand or by means of standard pocket calculators. There are various computer program packages that perform the analyses of

standard experiments (such as the ones discussed in this chapter). The *BMDP*, *SAS* and *SPSS* program series all have such packages. The advantage attached to the use of some of these packages is that they automatically test the underlying assumptions of a specific design and even guide the researcher as to what should be done if any of the assumptions are not valid. There are also programmable pocket calculators with which analysis of variance can be done within seconds.

Exercises

In all the problems below you may assume that the underlying assumptions for the applications of the respective techniques are met.

1 Because of petrol price hikes over the past few years, motor car petrol consumption has been subjected to strong scrutiny. Each car manufacturer claims that the petrol consumption of its own models is lower than that of other cars with the same cylinder capacity. In order to compare the products of four manufacturers, a car magazine decides to select five recently produced cars from each manufacturer randomly and to subject the twenty selected cars to identical road tests. The petrol consumption (ℓ/100 km) is given in the table below.

Manufacturer	Petrol consumption of 1 600 cm^3 cars				
Manufacturer 1	7.83	7.90	8.30	7.90	7.61
Manufacturer 2	8.60	8.90	8.15	9.10	9.45
Manufacturer 3	7.71	7.52	7.91	8.12	7.75
Manufacturer 4	9.02	9.01	8.71	8.80	8.99

Test at a 5% level of significance whether there are differences in the petrol consumption of the four 1 600cm^3 makes. If there are differences, use Scheffé's method of pairwise comparison to indicate all significant differences.

2 Repeat Example 14.2, but reduce each value in Table 14.6 by 50. Compare the f value obtained with the value obtained in Example 14.2. What do you conclude from this comparison?

3 The manager of a restaurant group that has branches in five cities wishes to determine whether the sales of a specific grill differ in the five cities. He instructs each branch manager to choose six days in the course of one month randomly and to note the number of grills served. The numbers obtained are as follows:

Cities	Number of grills sold					
Pretoria	61	39	45	21	18	53
Johannesburg	81	71	51	49	59	75
Cape Town	33	36	46	51	48	42
Durban	13	10	39	33	25	21
Port Elizabeth	33	27	16	14	21	25

Test at a 5% level of significance whether the average sales of this dish differ in the five cities.

4 The experimental design in Exercise 3 has the disadvantage that the experimental unit (days) is not homogeneous with regard to the characteristic observed. The few relatively large values noted at the Johannesburg restaurant were possibly all obtained on weekend days and the low figures at the Durban restaurant possibly reflected sales on poor business days (for example, Mondays). The assistant manager of the restaurant group recommends to the manager that every day of the week should be chosen by the branches randomly within the scope of one month and that the corresponding sales on these days should be observed. Execution of this randomized block design yields the following sales figures:

Cities	Day						
	Mo	Tu	We	Th	Fr	Sa	Su
Pretoria	25	28	36	44	61	65	52
Johannesburg	22	41	50	34	66	77	58
Cape Town	19	20	40	34	51	91	54
Durban	10	15	35	31	42	51	29
Port Elizabeth	14	21	40	24	55	50	43

Test again (with $\alpha = 0.05$) whether average sales in the five cities differ from one another. Would you say the inclusion of blocks (weekdays) was essential?

5 A large government subsidized corporation consists of three virtually identical plants. The government commissions a productivity investigation by the Department of Labour at each of these plants. The investigation is carried out by random selection of a number of workers at each plant. These workers are then subjected to intensive tests and questioning to determine their productivity. Since age affects productivity, the investigators decide to include the age factor as a block effect. To simplify matters, we shall consider the productivity (%) of fifteen workers only in the table below. Measurements are given as indices with 100 representing a normal productivity.

Plant	Age				
	[18, 20)	[20, 25)	[25, 35)	[35, 45)	[45, 65)
Plant 1	103	114	122	102	96
Plant 2	110	112	114	91	95
Plant 3	89	100	96	81	75

(a) Would you say that there is a difference in the mean productivity of workers in the three plants ($\alpha = 0.05$).

(b) Would you say that age affects worker productivity? ($\alpha = 0.05$).

6 The table below was compiled from Appendix A, Data Set 1 and represents an observed (4×3) factorial experiment with four observations per cell. The following method was used to compile the table: From each of the twelve subpopulations (four levels of Factor A and three levels of Factor B) four houses were chosen randomly and the house valuations and plot sizes were observed.

City (Factor A)	Valuations (R1 000) of 48 houses					
	Size of stand (Factor B)					
	1		2		3	
Durban	165.0	72.0	140.0	119.0	115.0	235.0
	157.0	84.0	120.0	90.0	94.0	220.0
Johannesburg	70.0	154.0	114.0	92.0	306.5	378.0
	117.0	213.0	249.0	147.0	402.0	445.0
Pretoria	86.0	64.5	121.0	293.0	138.0	135.0
	362.5	213.0	75.0	103.0	182.0	240.0
Cape Town	101.0	140.0	195.0	104.5	240.0	50.0
	110.0	104.0	280.0	68.0	185.0	47.9

(a) Test whether the factors A and B are significant ($\alpha = 0.05$).

(b) Is there any interaction between A and B? ($\alpha = 0.05$).

7 In a factorial experiment four wheat varieties and three fertilizers are of concern. To determine whether there are differences between the varieties and between the fertilizers, 36 homogeneous agricultural plots of equal size are assigned randomly to the twelve cells (three per cell). Before the planting season the soil is fertilized with the corresponding fertilizers after the harvest and the yield (per bag) per plot is determined.

Fertilizer (A)	Wheat yield of 36 agricultural plots							
	Variety (B)							
	1		2		3		4	
Type 1	46 19	35	41 11	26	35 34	21	26 12	15
Type 2	37 18	18	49 30	37	45 61	66	25 18	29
Type 3	32 32	40	65 74	76	66 67	60	45 41	39

(a) Test for the significance of factors A and B ($\alpha = 0.05$).

(b) Is there interaction between A and B? ($\alpha = 0.05$).

(c) If one or both factors are significant, use Scheffé's method of pairwise comparison to indicate all the significant level differences ($\alpha = 0.05$).

8 (a) The number of runs successively scored by four specialist batsmen in the South African cricket team in matches played during an overseas tour is as follows:

Batsman 1	10	0	53	121	70 n.o.	25	16	33	44		
Batsman 2	85	10	2	3	103	65 n.o.	32	22	143		
Batsman 3	3	2	6	33	68	32	16	21	0	4	2
Batsman 4	20	23	18	51	17	66					

(i) Test for differences in the batting ability of the batsmen ($\alpha = 0.05$).

(ii) Can the observed data really be regarded as the result of a completely randomized experiment? Motivate your answer.

(b) There are three types of matches played on tours, namely non-test matches, one-day internationals and five-day tests. A randomized block design has been used to compile the following table from the above information:

Batsman	Type of match		
	Non-test match	One day international	Five-day test
1	121	53	44
2	85	3	143
3	6	32	21
4	20	17	66

(i) Does this affect your opinion regarding the abilities of the four batsmen?

(ii) Are there differences between the three types of matches in respect of the performances of the batsmen ($\alpha = 0.05$)?

9 In order to compare the standard of teaching in high schools in a certain city, five standard 8 pupils are randomly chosen at each of ten high schools. The fifty pupils write an identical test in mathematics and the results are as follows:

School 1	35	60	52	54	48
School 2	93	80	71	54	66
School 3	38	52	60	72	70
School 4	50	52	71	43	66
School 5	82	71	65	72	75
School 6	33	24	62	65	70
School 7	91	30	70	80	65
School 8	62	66	71	65	63
School 9	53	62	70	73	57
School 10	74	76	66	68	85

(a) Test whether there is any difference between the standard of teaching in mathematics for standard 8 pupils at the ten schools ($\alpha = 0.05$).

(b) Can the experimental design be improved? If so, how?

10 Four pharmacists are asked to determine the percentage of methyl alcohol in a specific chemical compound. Each of them is given enough of the chemical compound to do three titrations. The table below shows the results. Are there significant differences in the findings of the pharmacists? ($\alpha = 0.05$)?

Pharmacist	Percentage of methyl alcohol		
1	64.99	64.04	64.38
2	65.15	65.13	64.88
3	64.72	64.48	65.16
4	64.20	64.10	64.55

11 An experiment is designed to compare the time required for each of four sequences for assembling a television set. Eight assemblers are used, each of whom assembles a television set according to each of the four methods, on four different days. The specific assembling sequence used by a particular assembler on a specific day is determined randomly. The assembling times (in minutes) are recorded in the accompanying table. Test at a 1% level of significance whether there is any difference in the times required for each of the four sequences. Do the assemblers differ ($\alpha = 0.05$)?

Sequences	Assemblers							
	1	2	3	4	5	6	7	8
A	63	58	68	72	75	69	74	66
B	49	51	49	59	64	52	66	54
C	60	57	63	71	68	72	78	67
D	65	61	60	76	71	69	81	63

If there are differences in assembling times, use Scheffé's method of pairwise comparison to indicate the significant differences in levels for the assemblers and the assembling methods ($\alpha = 0.01$).

12 The accompanying table shows the duration of both halves of matches in the Currie Cup rugby fixtures, chosen from the rugby schedule for a particular season. The selection was done randomly, but in such a way that four matches were refereed by Peter, four by Frank and four by Kevin.

Half	Referee											
	Peter				Frank				Kevin			
First	43	40	41	41	40	39	41	40	42	42	43	41
Second	45	42	41	43	40	41	40	41	45	46	42	47

(a) Does the duration of the first and second halves of Currie Cup matches differ ($\alpha = 0.05$)?

(b) Do the three referees differ in their approach to allowing injury time ($\alpha = 0.05$)?

(c) Is there any interaction between these two factors ($\alpha = 0.05$)?

13 A manufacturer of rubber products tests three chemical catalysts used to speed up the hardening process. The following table shows the hardening time (in minutes) in six experiments with each of the three catalysts.

Catalyst 1	12.5	12.2	12.8	12.9	12.4	13.3
Catalyst 2	9.9	10.3	9.8	9.6	10.9	10.0
Catalyst 3	12.1	11.9	11.8	12.1	12.0	11.7

(a) Test at a 1% level of significance whether the three catalysts differ.

(b) If the catalysts do differ, use Scheffé's method of pairwise comparison to indicate the significant differences.

14 The accompanying table shows the number of new motor vehicles registered in the RSA by year and type of vehicle. The information is

derived from *Statistical News Release P7151.1* issued by the Central Statistical Service.

Type of vehicle	Number of registrations			
	1989	1990	1991	1992
Cars	203 279	186 754	176 347	166 234
Minibuses	17 587	15 431	14 220	12 610
Buses	1 168	805	501	664
Commercial vehicles	77 690	73 362	64 015	62 293

Is there any significant difference in the sales of new vehicles over the four years ($\alpha = 0.05$)?

15 A dealer who sells Beretta shotguns receives complaints that the guns 'stick' when they are cocked. In the past, similar problems have been solved or reduced by polishing the working parts. A calibrated spring balance is used to measure the force of the parts used in the cocking process, after each shot. All these measurements are taken by the same individual. Eight different weapons and five kinds of ammunition are used by a marksman. In each case, five dry shots (a shot without ammunition) are fired with each weapon.

Weapon no.	Type of ammunition					
	1	2	3	4	5	Dry
1	8.5	5.0	4.5	4.5	5.5	4.0
	11.2	5.0	5.5	5.5	5.0	4.0
	8.0	5.0	7.0	4.0	5.5	4.5
	6.0	5.0	5.5	4.5	4.5	4.0
	8.0	6.5	4.5	4.5	5.0	4.0
2	5.5	5.5	5.5	6.0	5.5	4.5
	5.5	5.0	5.5	5.0	5.5	4.5
	5.5	5.5	6.0	5.5	8.5	4.5
	6.5	6.0	5.5	5.0	9.0	5.0
	5.5	5.0	7.0	6.5	5.5	4.0
3	7.0	7.5	4.5	5.0	7.5	4.5
	6.0	6.0	7.5	5.5	6.0	4.5
	9.0	5.0	6.5	5.5	5.5	4.0
	8.5	7.5	5.5	6.5	5.5	4.5
	8.0	5.5	5.0	6.0	8.5	4.5
4	5.0	7.5	5.5	5.5	6.5	6.0
	8.0	5.5	5.0	5.5	5.0	5.0
	8.0	5.5	5.5	5.0	7.0	5.0
	5.0	10.0	5.0	5.0	5.5	5.0
	7.5	5.0	6.5	5.0	5.0	5.5

Weapon no.	Type of ammunition					
	1	2	3	4	5	Dry
5	8.0	6.5	5.0	5.5	5.0	4.5
	5.0	5.0	7.0	8.0	5.5	5.0
	6.0	5.5	5.5	9.0	5.0	4.5
	7.5	7.5	5.5	5.0	4.5	4.5
	3.0	5.5	6.5	4.5	7.0	4.5
6	4.5	5.0	4.0	6.0	5.5	4.5
	5.0	5.5	4.0	4.0	3.5	4.5
	6.0	5.0	4.5	5.0	4.5	4.5
	5.5	5.0	5.0	4.5	5.5	4.5
	5.5	5.0	4.5	5.0	5.0	4.5
7	13.0	13.0	11.5	5.0	8.5	5.0
	3.0	6.0	13.0	7.5	5.5	5.0
	13.0	8.0	8.5	7.0	5.0	5.0
	13.0	13.0	13.0	7.0	8.5	5.0
	13.0	9.5	5.0	5.0	6.5	4.5
8	7.9	5.5	7.0	5.5	5.0	4.5
	9.0	4.5	7.5	8.5	4.5	4.5
	7.5	6.0	5.0	6.0	4.5	5.0
	7.5	6.0	5.5	5.5	4.5	5.0
	7.5	5.5	5.5	10.0	5.0	5.0

Data obtained from MBL1 students J A J de Necker and E P van Eeden

(a) Is there a difference in the cocking force of the respective weapons ($\alpha = 0.05$)?

(b) Is there a difference in the cocking forces with different kinds of ammunition ($\alpha = 0.05$)?

(c) Is there any significant interaction between the weapon and the kind of ammunition used ($\alpha = 0.05$)?

16 In the course of an agricultural experiment steers are given different rations. Important criteria in measuring growth are fodder turnover and average daily increase in mass. The steers are randomly divided into four groups and given the following rations:

A: constant energy level and protein level

B: constant energy level and decreased protein level

C: increased energy level and decreased protein level

D: increased energy level and constant protein level

The data is as follows:

Steer number	Starting mass (kg)	Final mass (kg)	Number of feeding days	Cumulative consumption (kg)
Treatment A				
101	279	403	78	682
102	282	431	78	733
103	273	412	85	689
104	258	420	85	773
105	225	415	125	1 022
106	212	438	146	1 182
107	215	334	78	704
108	192	411	146	1 203
109	185	380	127	894
110	198	375	127	855
111	181	395	146	1 116
Treatment B				
201	280	435	78	717
202	270	431	127	977
203	260	399	85	811
204	240	403	127	1 033
205	218	421	127	1 046
206	233	408	127	1 024
207	223	379	127	1 026
208	213	416	146	844
209	206	405	146	949
210	198	408	146	1 012
Treatment C				
301	262	402	78	714
302	257	413	78	724
303	260	416	85	810
304	271	416	85	782
305	211	401	127	1 092
306	220	395	146	1 025
307	212	435	146	998
308	210	421	127	1 107
309	198	405	146	977
310	193	418	146	1 204
311	189	370	146	1 221
Treatment D				
401	308	455	78	677
402	267	418	85	640
403	266	401	85	767
404	247	433	127	1 032
405	225	380	127	966
406	238	423	127	966
407	216	430	146	1 120
408	208	413	127	1 024
409	192	410	146	986
410	187	365	146	756
411	189	346	127	826

Data provided by HPF du Preez, Langmoed Stud, Coligny

(a) Calculate for each treatment:

$$\text{Fodder turnover} = \frac{\text{cumulative kg of fodder consumed}}{\text{kg increase in mass}}$$

(b) Compile a factorial design in a table as shown below using fodder turnover as the criterion.

Protein level	Energy levels	
	Constant	Increased
Constant	Treatment A	Treatment D
Decreased	Treatment B	Treatment C

(c) Test for the significance of protein and energy levels ($\alpha = 0.05$).

(d) Is there any interaction between the protein and energy levels ($\alpha = 0.05$)?

(e) If one or both factors are found to be significant, apply Scheffé's method of pairwise comparison to indicate all the significant levels ($\alpha = 0.05$).

17 Repeat Exercise 16, but use the variable ADI (average daily increase) with

$$\text{ADI} = \frac{\text{kg increase}}{\text{number of feeding days}}$$

18 A medical researcher has collected data on the stress levels at which adrenalin is released after injection with a certain drug. The aim of the investigation was to test the effect of the drug in inhibiting the release of adrenalin. A sample of fifty males was selected from volunteers. The sample was randomly divided into five groups. Each group comprised ten male individuals. The first was a control group and was injected with a placebo. The second group was injected with 10 cc, the third, fourth and fifth groups were injected with 20, 30 and 40 cc respectively. Each patient was then subjected to a stress situation and the stress level at which adrenalin is released was measured in each individual.

Stress levels at which adrenalin was released

Placebo		10 cc		20 cc		30 cc		40 cc	
14	15	22	16	18	23	22	28	27	23
19	23	13	21	22	25	19	22	26	24
16	22	18	25	19	21	21	20	22	29
21	26	17	18	16	17	28	23	26	27
18	17	19	23	22	25	24	24	25	22

Data obtained from MBL1 group, Umtata 00193

(a) Test the null hypothesis that there is no significant difference in the stress level at which adrenalin is released for the five treatment groups ($\alpha = 0.05$).

(b) Apply Scheffé's method of pairwise comparison to indicate any significant differences in levels if there are any.

19 The effectiveness of three methods of treatment for burns was tested in a mine hospital. The patients in the investigation were male employees with burns admitted to the hospital over a nine-week period. For the first three weeks, patients were treated according to method 1. During weeks four to six they were treated according to method 2 and during the last three weeks a third method of treatment was applied. At all times, patients were informed of the nature and extent of the investigation. Any patient who was unwilling to participate in the experiment was excluded from the investigation. The patients in each group were comparable in terms of age, size and extent of burns. The respective treatments for the burns were applied on admission and repeated on a daily basis until healing had been effected. The array for the number of days that elapsed before patients from each of the three groups had healed is reflected below.

Array of number of days for wounds to heal

Method 1	Method 2	Method 3
9	5	6
12	5	7
15	6	8
15	6	9
18	6	10
22	7	10
24	8	12
24	8	16
26	9	19
31	10	21
33	11	23
	13	23
	15	24
	16	28
	17	31
	24	32
	26	38
	33	

Data obtained from MBL1 group, Rustenburg 00193

(a) Test at a 1% level of significance whether the mean number of days required for healing varied for the three methods of treatment.

(b) Are you of the opinion that the experimental design was correctly executed? How could it be improved?

20 A sugar company investigates the financial viability of increasing the amount of sucrose extracted per hectare by applying chemical ripeners to sugar cane. Four methods are applied:

(i) no ripeners (control);
(ii) chemical A;
(iii) chemical B;
(iv) chemicals A and B.

An experiment was conducted on a large sugar cane field that had been divided into 24 blocks. Six of the blocks were randomly treated with chemical A, six with chemical B and six with A and B. A control section of six blocks was left untreated. Eleven samples were harvested from each block so that a total of precisely $198m^2$ of cane was harvested for testing purposes. The sugar cane was weighed to determine the yield per hectare in ton and samples of the cane were taken and subjected to laboratory tests to determine the sucrose level. The data for the 24 blocks is as follows:

Control		Chemical A		Chemical B		Both A and B	
Yield (Tons/ha)	Sucrose level	Yield (Tons/ha)	Sucrose level	Yield (Tons/ha)	Sucrose level	Yield (Tons/ha)	Sucrose level
79	16.09	77	16.72	75	18.62	78	18.73
76	14.09	94	17.36	95	17.30	93	18.86
101	16.17	82	17.93	95	17.80	98	19.01
96	15.99	98	18.38	99	18.18	93	16.46
89	17.87	101	17.26	91	18.50	86	18.63
80	16.55	76	17.17	79	18.82	80	18.33

Data obtained from MBL1 group, Swaziland 00193

(a) Test at a 5% level of significance whether the mean sucrose levels are equal for the four treatments.

(b) Apply Scheffé's pairwise comparison technique to indicate any significant differences in the levels.

21 Consider the harvest in the sugar experiment referred to in Exercise 20.

(a) Test at a 5% level of significance whether the yield (ton/ha) is the same for each of the four methods of treatment.

(b) Apply Scheffé's pairwise comparison technique to indicate any significant differences in the levels.

CHAPTER 15

Contingency tables

1 Introduction

In Chapter 3 we saw that a frequency distribution for a quantitative or qualitative univariate data set offers a simple but important representation. We can also regard a frequency distribution as a **one-way contingency table**, since each observation is categorized, according to realization (or contingency), into one of a number of mutually exclusive 'classes of events'. Often information is obtained on two or more characteristics of n individuals or entities. Measurement of such characteristics then yields multivariate data sets. Some of the characteristics can be measured quantitatively, while others are measured qualitatively. Such multivariate data sets can be represented simply and meaningfully in **multi-way contingency tables**.

A **contingency table** is a summary of a univariate or multivariate data set in the form of either a one-dimensional (one-way) or a multi-dimensional (multi-way) frequency distribution.

Tables 15.1 to 15.4 are typical examples of contingency tables. The variables or attributes in these contingency tables are often called 'factors' or characteristics. The single factor in Table 15.1 is a discrete variable, viz. 'outcome of a dice roll'.

Table 15.1 One-way contingency table of the number of ones, twos, ..., sixes obtained by 160 rolls of a dice

One	Two	Three	Four	Five	Six	Total
18	30	32	22	24	34	160

Table 15.2 Two-way contingency table of the success of two types of medical treatment to which 143 patients suffering from the same illness are subjected for two consecutive weeks

Treatment A	Treatment B		Total
	Favourable	Unfavourable	
Favourable	59	45	104
Unfavourable	17	22	39
Total	76	67	143

Table 15.3 Two-way contingency table of a random sample of 893 houses classified according to age and number of bedrooms

Bedrooms	Built before 1960	Built 1960–1969	Built 1970–1979	Built 1980–1989	Built 1990 or later	Total
2	41	11	5	9	7	73
3	135	70	92	169	167	633
4 and more	38	17	37	48	47	187
Total	214	98	134	226	221	893

Table 15.4 Three-way contingency table of the number of unemployed black and coloured workers (in thousands) during May 1980, subdivided according to age group and gender

Age group	Population group			
	Coloureds		Blacks	
	Men	Women	Men	Women
[16, 20)	8	10	43	34
[20, 30)	8	12	99	109
[30, 40)	2	7	32	61
[40, 50)	3	4	25	40
[50, 65)	3	2	23	20

Source: *Statistical News Releases P 27.3* and *P 27.4*, May 1980

The above contingency tables clearly show that univariate and multivariate data sets can be tabulated meaningfully, despite the fact that the nature of the characteristics may differ.

It is often useful to identify a contingency table according to the number of factors and also the number of categories (levels) per factor. Thus the above contingency tables can also be identified as follows:

Table 15.1: (1×6) contingency table
Table 15.2: (2×2) contingency table
Table 15.3: (3×5) contingency table
Table 15.4: $(5 \times 2 \times 2)$ contingency table

The first figure indicates the number of row categories or factor A categories and the second figure the number of columns or factor B categories. Note that Table 15.4 is a two-dimensional representation of a three-way table, in which factors A, B and C are respectively 'age group', 'population group' and 'gender'.

Although the description of a univariate or multivariate raw data set already provides valuable information when represented in a contingency table, it is usually only the first step in a statistical analysis of the data. The purpose is to draw certain justified conclusions about the population from which the data set is derived. In this regard the χ^2 distribution plays an important role. This distribution is used in virtually all the procedures for testing hypotheses discussed in this chapter.

2 The χ^2 goodness-of-fit test

The assumption that a sample of observations is taken from a normal population forms the basis for many of the inference techniques discussed in Chapters 11, 12, 13 and 14. If this assumption is not justified, it is important that the researcher should be aware of this fact, so that alternative inference techniques (such as the ones in Chapter 16) can be used. In this section we indicate how we can test whether or not an observed univariate data set does come from a normal population. One test that can be used here is known as the χ^2 goodness-of-fit test (abbreviated as the χ^2 fit test).

> If a relatively large number of observations of a stochastic variable are represented in a one-way contingency table, the **χ^2 goodness-of-fit test** is used to decide whether the frequency distribution can be reconciled with an assumed theoretical distribution.

Although this test is used mainly to confirm the normality of a data set, it can be used to confirm any underlying probability model. The following simple example explains the procedure.

EXAMPLE 15.1

Table 15.1 is a summary of the outcomes of 160 rolls of a dice. Suppose we want to test whether the dice is balanced. If X indicates the outcome of a roll of the dice, we then want to test whether the probability distribution of X is given by

$$p(x) = \tfrac{1}{6};\ x = 1, 2, 3, 4, 5, 6.$$

On the assumption that X does in fact have this probability distribution, the theoretical or expected frequency of ones in 160 rolls is equal to $\frac{160}{6}$. The same theoretical frequency applies to the other possible outcomes. Table 15.5 now shows both the observed frequencies o and the expected frequencies e. The last row contains the values $\frac{(o-e)^2}{e}$.

Table 15.5 Calculation table for the χ^2 goodness-of-fit test for use in deciding whether or not a dice is balanced

	One	Two	Three	Four	Five	Six	Total
o_i	18	30	32	22	24	34	160
e_i	26.67	26.67	26.67	26.67	26.67	26.67	160.02
$\frac{(o_i - e_i)^2}{e_i}$	2.8185	0.4158	1.0652	0.8177	0.2673	2.0146	7.3991

If we now state the null hypothesis that the dice is balanced, namely that

$$H_0\text{: } p(x) = \tfrac{1}{6};\ x = 1, 2, \ldots, 6,$$

it can be shown that under H_0

$$\sum_{i=1}^{6} \frac{(o_i - e_i)^2}{e_i} = 7.3991$$

can be regarded as an observed value from a $\chi^2(5)$ distribution. The alternative here is that the dice is unbalanced. Large differences between corresponding o and e values will lead to a large value for $\sum_{i=1}^{6} \frac{(o_i - e_i)^2}{e_i}$. Such differences are then an indication of a poor correspondence between the observed frequencies and the theoretical frequencies. In such cases the null hypothesis is rejected. From Table B.4 $\chi^2_{5,\,0.05} = 11.070$ and, since the value 7.3991 is smaller than 11.070, it follows that H_0 cannot be rejected at a 5% level of significance. We therefore assume that the dice is balanced.

In the above example the distribution of the relevant stochastic variable was fully specified under H_0. Often the null hypothesis does not specify the distribution fully and H_0 can, for example, merely read: 'The observations are taken from a normal population.' In such cases some or all of the parameters of the distribution should then be estimated, for instance μ and σ^2 in the latter case of null hypothesis. If we indicate the number of parameters to be estimated by p, the number of degrees of freedom of the distribution of the test statistic should be reduced by p. See Table 15.6 in this regard. This table describes the application of the χ^2 goodness-of-fit test on a $(1 \times k)$ observed frequency distribution. In Example 15.2 the application of this procedure is illustrated in the case of observations that are presumably taken from a normal population.

Table 15.6 χ^2 goodness-of-fit test for an observed ($1 \times k$) contingency table

Test statistic	Distribution of test statistic under H_0	Table with critical values
$C = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$	n large: $\chi^2(k-p-1)$ provided $E_i \geq 5$ for all i	Table B.4
Hypotheses	**Decision rule: Reject H_0 if**	***p*-value**
H_0: Observed data are observations of a stochastic variable with specified theoretical distribution H_1: H_0 is untrue	$c \geq \chi^2_{k-p-1,\,\alpha}$	$p = P(C \geq c)$

EXAMPLE 15.2

In Chapter 10 we assumed that the price per square metre of houses is normally distributed. Based on the price per square metre of the random sample of 100 Johannesburg houses of which the frequency distribution is given in Table 3.7, we now show how the χ^2 goodness-of-fit test can be used to arrive at a decision in this regard.

Table 15.7 shows how the value of C can be calculated. The e_i values reflected in this table are calculated as follows: The arithmetic mean and standard deviation of the grouped data are

$$\bar{x} = 914 \text{ rand} \quad \text{and} \quad s = 341.423 \text{ rand}.$$

We know that $\bar{X}$ and S^2 are unbiased estimators for the mean μ and the variance σ^2 of the distribution of X. If we now assume that the observed sample of the price per square metre is obtained from a n(914, $(314.423)^2$) population, we can calculate the theoretical frequency of, for instance, the first interval, namely [100, 300) as follows:

$$\begin{aligned} P(100 \leq X < 300) &= P\left(\frac{100 - 914}{341.423} \leq Z < \frac{300 - 914}{341.423}\right) \\ &= P(-2.38414 \leq Z < -1.79836) \\ &= H(2.38414) - H(1.79836) \\ &\approx 0.0275 \qquad \text{(see Table B.2)} \end{aligned}$$

Consequently, we can estimate that the expected number of values in the first class would be

$$\begin{aligned} e_i &= n \times P(100 \leq X < 300) \\ &= 100 \times 0.0275 \\ &= 2.75 \end{aligned}$$

if the sample is drawn from this normal population. The other estimated frequencies are calculated similarly.

Table 15.7 Calculation of the test statistic for testing the null hypothesis that 100 observed prices per square metre are derived from a normal population

Class interval	Observed frequency o_i		Expected frequency e_i		$\frac{(o_i - e_i)^2}{e_i}$
[100, 300)	2	10	2.75	10.41	0.0161
[300, 500)	8		7.66		
[500, 700)	18		15.28		0.4842
[700, 900)	25		21.83		0.4603
[900, 1 100)	19		22.34		0.4994
[1 100, 1 300)	12		16.38		1.1712
[1 300, 1 500)	10	16	8.61	11.85	1.4534
[1 500, 1 700)	6		3.24		
	100		98.09		$c = 4.0846$

In Table 15.7 adjacent classes are grouped together to give frequencies of at least 5. The steps to follow for testing the hypothesis are:

Step 1

H_0: The data set is obtained from a normal population.
H_1: The data set is not obtained from a normal population.

Step 2

Choose $\alpha = 0.05$. Since $k = 6$ (after the necessary grouping of classes) and $p = 2$, it follows that under H_0

$$C = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$$

is approximately distributed as $\chi^2(3)$.

Step 3

From Table B.4, $\chi^2_{3,\,0.05} = 7.815$ so that we reject H_0 if $c \geq 7.815$ or if the p-value ≤ 0.05.

Step 4

According to Table 15.7, $c = 4.0846$.

Step 5

Since $c < 7.815$, it follows that H_0 cannot be rejected at a 5% level of significance. The p-value = 0.25247. The assumption that this sample of prices per square

metre of Johannesburg houses has been obtained from a normal population, is thus verified by the χ^2 goodness-of-fit test.

Notes

(a) Although the χ^2 goodness-of-fit test can be used to decide whether an observed data set is taken from any specified distribution, it is not always easy to calculate the expected frequencies, since comprehensive tables for each probability distribution are not available.

(b) The χ^2 approximation of the test statistic for the goodness-of-fit test is, strictly speaking, applicable only if the number of sample observations is large and all the expected frequencies are at least 5. However, some writers maintain that the latter requirement is unnecessarily strict and that e values as low as 2 or 3 are still acceptable.

(c) A quick procedure for examining a large observed data set for normality is to use so-called **normal probability paper**. The relative cumulative frequency polygon of the data set is plotted on this special graph paper (often with omission of the first and last points). The scale of the vertical axis on the graph paper is such that the frequency polygon will approximately represent a straight line if the data set has been taken from a normal population. No hypothesis testing is involved in this technique, so it is, to some extent, a subjective method.

(d) For small data sets ($n \leq 50$) the normality of a data set can be confirmed by means of the so-called **rankits procedure**. This is also a graphic method.

(e) The **Kolmogorov-Smirnov test** is a well-known goodness-of-fit test that is useful for small samples.

3 The (2 × 2) table (fourfold table)

With reference to Table 15.2, two questions are now important:

(i) Are a patient's reactions to the two medical treatments independent of one another? In other words, will a patient's favourable or unfavourable reaction to treatment A influence his reaction to treatment B?

(ii) Is the ratio of patients who react favourably to treatment A equal to the ratio of those who reacted favourably to treatment B?

The test used to answer question (i) is known as the **(2 × 2) independence test**, while **McNemar's** test is used to obtain an answer to question (ii).

3.1 The (2 × 2) independence test

The general representation of a (2 × 2) contingency table for two factors A and B is given as in Table 15.8.

Table 15.8 General representation of an observed (2 × 2) contingency table

A (row factor)	B (column factor) B_1	B_2	Total
A_1	o_{11}	o_{12}	r_1
A_2	o_{21}	o_{22}	r_2
Total	k_1	k_2	n

If n entities are classified simultaneously according to either the presence or the absence of two characteristics or qualities, the **(2 × 2) independence test** is used to determine whether or not there is dependence (interaction) between the characteristics.

In terms of probability theory, independence of the row and column factors (see Table 15.8) would imply that

$$P(A_i \times B_j) = P(A_i)\ P(B_j);\ i, j = 1, 2$$

where A_i is the event that a randomly selected entity is placed in the ith row and B_j the event that a randomly selected entity is placed in the jth column. The test procedure explained in Table 15.9 was proposed by F. Yates and is illustrated in Example 15.3.

Table 15.9 (2 × 2) independence test

Test statistic	**Distribution of test statistic under H_0**	**Table with critical values**
$Y = \dfrac{n\left(\lvert O_{11}O_{22} - O_{12}O_{21}\rvert - \dfrac{n}{2}\right)^2}{R_1R_2K_1K_2}$	approximately $\chi^2(1)$ provided n is large and $\dfrac{R_iK_j}{n} \geq 5$ for all i and j	Table B.4
Hypotheses	**Decision rule: Reject H_0 if**	**p-value**
H_0: Row and column factors independent H_1: Row and column factors dependent	$y \geq \chi^2_{1,\alpha}$	$p = P(Y \geq y)$

EXAMPLE 15.3

Table 15.10 was obtained by drawing a random sample of 107 residences on the West Rand and classifying them according to age and property valuation. Normally

one would expect the age and property valuation to display a negative relation, that is, an old house would be worth significantly less than a new house. On the other hand, old houses are usually located more centrally, which has a positive effect on the property valuation. The nett effect of age is therefore difficult to predict. We shall now test for independence of these two factors.

Table 5.10 (2 × 2) observed table of 107 residences on the West Rand, classified according to age and property valuation

Property valuation	Age		Total
	Built before 1980	Built 1980 and later	
R150 000 and less	22	29	51
More than R150 000	13	43	56
Total	35	72	107

Step 1

H_0: The age and property valuation of a West Rand house are independent
H_1: The two factors are dependent.

Step 2

Choose $\alpha = 0.05$. Under H_0

$$Y = \frac{107\left(\left|O_{11}O_{22} - O_{12}O_{21}\right| - \frac{107}{2}\right)^2}{R_1R_2K_1K_2}$$

is distributed approximately as $\chi^2(1)$.

Step 3

Reject H_0 if $y \geq \chi^2_{1,0.05} = 3.842$ or if the p-value ≤ 0.05.

Step 4

From Table 15.10 it follows that

$$y = \frac{107\left(\left|22 \times 43 - 13 \times 29\right| - \frac{107}{2}\right)^2}{51 \times 56 \times 35 \times 72}$$
$$\approx 3.9508$$

Step 5

Since $y \geq 3.842$, H_0 is rejected at a significance level of 5%. The p-value = 0.04685. It therefore becomes evident that the age of a residence on the West Rand does affect its property valuation.

Notes

(a) In cases where n is small or where the expected frequencies under H_0 are not at least 5, **Fisher's exact test** for a (2×2) contingency table can be used.

(b) In Example 15.3 the fact that $o_{11}o_{22}$ is considerably larger than $o_{21}o_{12}$ indicates that the dependence between age and property valuation lies in a specific direction, namely that newer residences apparently have a higher property valuation than older residences. This is known as a positive relationship. If the null hypothesis of independence is tested against such a one-sided alternative of positive dependence, the alternative can be accepted at a $100\alpha\%$ level of significance, provided that $y \geq \chi^2_{1,\,2\alpha}$. Thus in Example 15.3 such a positive relationship is found to exist when testing is done at a 5% level of significance. In general, the null hypothesis of independence can also be tested against an alternative of negative dependence (if the problem suggests such an alternative and $o_{21}o_{12}$ is much greater than $o_{11}o_{22}$). This alternative is also accepted if $y \geq \chi^2_{1,2\alpha}$

(c) Based on a (2×2) table, the ϕ **coefficient** can be used to represent the degree and direction of dependence between the two characteristics A and B. The coefficient is defined as

$$\phi = \frac{o_{11}o_{22} - o_{21}o_{12}}{\sqrt{r_1 r_2 k_1 k_2}}$$

For the data in Table 15.10 we find that

$$\phi = \frac{946 - 377}{\sqrt{51 \times 56 \times 35 \times 72}}$$
$$= 0.21$$

Although there is a positive dependence between age and property valuation for West Rand residences, the relation is not very strong, since it can be shown that $-1 \leq \phi \leq 1$, in which the absolute size of ϕ reflects the strength of the relationship. When calculating the ϕ coefficient or testing H_0 against a one-sided alternative, it is important to remember that the sign or direction of the dependence is determined by the arrangement of the classes A and B.

3.2 McNemar's test for two equally probable characteristics

This test is applied in, for example, clinical experiments in which n patients are subjected to medications A and B and the reaction of each patient is recorded as favourable or unfavourable. For the doctor it is then important to test whether the probabilities of a patient reacting favourably to the two medications are the same.

The McNemar test can also be used to analyze the shift towards or away from a certain political party by asking n voters during, say, two consecutive elections, to indicate their support of or opposition to the party. The test is also applied in the behavioural sciences where the incidence of certain physical or behavioural characteristics of individuals is investigated.

If n entities are classified simultaneously according to either the presence or absence of two characteristics or qualities, the **McNemar test** is used to determine whether or not the two characteristics occur with equal probability in the population entities concerned.

The presence or absence of a certain characteristic is indicated by A and $\bar{A}$ respectively. B and $\bar{B}$ have the same meaning for a second characteristic. Now consider the observed frequencies in a random sample of size n, as indicated in Table 15.11. In this case we test to see if

$$P(A) = P(B)$$

Table 15.11 Representation of an observed (2 × 2) contingency table for applying the McNemar test

First characteristic	Second characteristic B	Second characteristic $\bar{B}$	Total
A	o_{11}	o_{12}	r_1
$\bar{A}$	o_{21}	o_{12}	r_2
Total	k_1	k_2	n

Table 15.12 gives a summary of the test procedure and Example 15.4 explains its application.

Table 15.12 McNemar's test for two equally probable characteristics

Test statistic	Distribution of test statistic under H_0	Table with critical values
$M = \dfrac{(O_{12} - O_{21})^2}{O_{12} + O_{21}}$	n large: approximately $\chi^2(1)$ provided $O_{12} + O_{21} \geq 10$	Table B.4
Hypotheses	**Decision rule: Reject H_0 if**	***p*-values**
H_0: $P(A) = P(B)$ H_1: $P(A) \neq P(B)$	$m \geq \chi^2_{1,\alpha}$	$p = P(M \geq m)$

EXAMPLE 15.4

With reference to Table 15.2, it may be necessary to test whether patients will react favourably to treatment A and treatment B with the same measure of probability. The two events are represented by A and B.

Step 1

H_0: $P(A) = P(B)$
H_1: $P(A) \neq P(B)$

Step 2

Choose $\alpha = 0.01$. If H_0 is true, then

$$M = \frac{(O_{12} - O_{21})^2}{O_{12} + O_{21}}$$

is approximately distributed as $\chi^2(1)$.

Step 3

Reject H_0 if $m \geq \chi^2_{1,0.01} = 6.635$ or if the p value ≤ 0.01.

Step 4

From Table 15.2 it follows that

$$m = \frac{(45 - 17)^2}{45 + 17}$$
$$= 12.645$$

Step 5

Since $m \geq 6.635$, it follows that H_0 is rejected at a 1% level of significance. The p-value = 0.00038. The probability of favourable reactions to the two treatments thus differ significantly.

Notes

(a) If n is small, we can use the fact that o_{12} can be regarded as an observation of a bi$(o_{12} + o_{21}; \pi)$ distribution. In this case the null hypothesis reduces to H_0: $\pi = \frac{1}{2}$ and the binomial tables can be used to test H_0.

(b) If n entities are classified according to either the presence or the absence of $\ell > 2$ characteristics or qualities, a 2^ℓ table is formed. Under certain assumptions **Cochran's test** can be used to establish whether all these characteristics occur with equal probability in an entity or individual. In Chapter 16, Section 4.2, Note (c), a simple representation of such a situation is discussed briefly.

4 The $(I \times J)$ table

4.1 The $(I \times J)$ independence test

In two-way contingency tables where one or both factors have more than two categories, the independence hypothesis is very important. In Table 15.3, 893 residences in the Johannesburg area are classified according to number of bedrooms and age. Are these two factors independent or does the decade in which the house was built perhaps have some bearing on the size of the house and therefore also on the number of bedrooms in the house? To the geneticist, for example, it is important to determine whether there is a relation between two genetic characteristics in a living organism in the sense that the category into which a specific organism is classified in terms of one of its characteristics also has a bearing on the classification of the organism in terms of a further characteristic. If a hundred women are classified according to the colour of their eyes and the 'true' colour of their hair, we should like to test the validity of the generally accepted opinion that these two characteristics are dependent. Dependence in this case would indicate that a woman with, say, brown eyes would have brown or black hair rather than blond or red hair.

The independence test for an $(I \times J)$ table is based on the general representation of this contingency table as set out in Table 15.13.

Table 15.13 General representation of an observed $(I \times J)$ contingency table

	B (column factor)				
A (row factor)	B_1	B_2	...	B_J	Total
A_1	o_{11}	o_{12}	...	o_{1J}	r_1
A_2	o_{21}	o_{22}	...	o_{2J}	r_2
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
A_I	o_{I1}	o_{I2}	...	o_{IJ}	r_I
Total	k_1	k_2	...	k_J	n

In terms of probability theory, the independence of factors A and B means that for each i and each j

$$P(A_i \cap B_j) = P(A_i)\,P(B_j)$$

where A_i is the event that an entity is classified in row A_i and B_j is the event that an entity is classified in column B_j. The test procedure is explained in Table 15.14 and illustrated in Example 15.5.

Table 15.14 The ($I \times J$) independence test

Test statistic	Distrubution of test statistic under H_0	Table with critical values
$C = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ with $E_{ij} = \frac{R_i K_j}{n}$	approximately $\chi^2\{(I-1)(J-1)\}$ if n is large and $E_{ij} \geq 5$ for all i and j	Tabel B.4
Hypotheses	**Decision rule: Reject H_0 if**	**p-value**
H_0: Row and column factors independent H_1: Row and column factors dependent	$c \geq \chi^2_{(I-1)(J-1),\,\alpha}$	$p = P(C \geq c)$

EXAMPLE 15.5

Here we consider the observed (3×5) contingency table set out in Table 15.3. Table 15.15 gives the observed o and the calculated e values. The calculated e values are indicated in brackets and are the estimates of the expected frequencies under H_0. The e values are calculated according to the following formula:

$$e_{ij} = \frac{r_i \times k_j}{n}$$

and e_{11} is for example calculated as follows:

$$e_{11} = \frac{r_1 \times k_1}{n}$$
$$= \frac{(73 \times 214)}{893}$$
$$= 17.49$$

Table 15.15 Observed and estimated cell frequencies in testing for independence between the age and number of bedrooms, based on a random sample of 893 residences in Johannesburg

A: Number of bedrooms	*B*: Age of house					Total
	Built before 1960	Built 1960–1969	Built 1970–1979	Built 1980–1989	Built 1990 and later	
2	41 (17.49)	11 (8.01)	5 (10.95)	9 (18.47)	7 (18.07)	73
3	135 (151.69)	70 (69.47)	92 (94.99)	169 (160.20)	167 (156.66)	633
4 and more	38 (44.81)	17 (20.52)	37 (28.06)	48 (47.33)	47 (46.28)	187
Total	214	98	134	226	221	893

The steps to follow for testing the hypothesis are:

Step 1

H_0: The age of a residence and the number of bedrooms in the house are independent.
H_1: The age of a residence does influence the number of bedrooms in the residence.

Step 2

Choose $\alpha = 0.01$. Since $(I - 1)(J - 1) = 8$, it follows that under H_0

$$C = \sum_{i=1}^{3} \sum_{j=1}^{5} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

is approximately distributed according to the $\chi^2(8)$ distribution.

Step 3

$\chi^2_{8,\,0.01} = 20.090$ so that H_0 is rejected if $g \geq 20.090$ or if the p-value is ≤ 0.01.

Step 4

Using the o and e values in Table 15.4, it follows that

$$c = \frac{(41 - 17.49)^2}{17.49} + \frac{(11 - 8.01)^2}{8.01} + \ldots + \frac{(47 - 46.28)^2}{46.28}$$
$$\approx 55.20$$

Step 5

Since $c \geq 20.090$, it follows that H_0 is rejected at a 1% level of significance. The p-value = 0.00000000404. There is therefore no doubt that the age of the residence and the number of bedrooms are dependent.

Notes

(a) A problem that often arises when $(I \times J)$ contingency tables are analyzed is that the estimated expected frequencies are lower than the required 5. If this is the case, and one or both factors have more than two categories, two or more adjacent rows and/or columns can be combined to create cells that meet this requirement. Such a combination is meaningful only if the factors involved are ordinal. This combination does, however, have the serious disadvantage that a substantial amount of information is lost and it should therefore be used only where absolutely essential. As mentioned in Note (b) of Section 15.2, many authors regard the requirement that all E_{ij}s be at least 5 as unnecessarily stringent.

(b) A general measure of the degree of relation between the factors A and B in the $(I \times J)$ table is the so-called **contingency coefficient**

$$cc = \sqrt{\frac{c}{c + n}}$$

in which c is the calculated $(I \times J)$ independence test statistic. From the definition of cc it follows that this coefficient is always positive and can therefore not be used to indicate the direction (if meaningful) of the relationship. Since the maximum value of cc is equal to $\sqrt{(m - 1)/m}$, with m the minimum of I and J, it is meaningful to compare only the cc values from tables with the same m value. For the data in Table 15.15 it follows that

$$cc = \sqrt{\frac{55.20}{55.20 + 893}}$$
$$= 0.24$$

This value is considerably higher than $cc = 0.130$ which can be found for Table 15.16, but still relatively low if we keep in mind that, for these two tables, the maximum value of cc is 0.82.

4.2 Test for the homogeneity of independent samples

Thus far we have assumed throughout that a two-way contingency table develops because **one sample of n entities** has been classified according to two characteristics. In these cases, when the contingency table was drawn up, we had no knowledge concerning the row and column totals prior to completion of sampling. It is, however, also possible to draw up an $(I \times J)$ contingency table by classifying each of J independent samples of entities according to a single characteristic with I categories. Of course, the sample sizes are known, and so the column totals of the $(I \times J)$ contingency table are known beforehand. If the $(I \times J)$ independence test is now applied to the contingency table thus formed and we find that the row factor and the 'column factor' (with the different samples as the columns) are independent of each other, it actually means that the J underlying populations do not differ from one another.

The **test for the homogeneity of J independent samples** of entities, each classified according to a single characteristic with I categories, determines whether or not these J samples are taken from populations that have the same distribution.

Since this test is actually an independence test and has already been discussed in Section 4.1, the test procedure is not given in detail. The following example will thus suffice.

EXAMPLE 15.6

To investigate whether or not the East Rand, West Rand and Johannesburg differ with regard to the number of bedrooms per residence, three random samples of size 100 houses each are taken from the three areas and each house is classified according to the number of bedrooms. The results obtained are summarized in Table 15.16 where the expected cell frequencies are again indicated in brackets. The steps to follow for testing the hypothesis are:

Table 15.16 Observed (3 × 3) table of 300 residences classified according to area and number of bedrooms. Estimated cell frequencies are also indicated

A: Number of bedrooms	*B*: Area (sample) East Rand	West Rand	Johannesburg	Total
2	6 (7)	7 (7)	8 (7)	21
3	76 (76.67)	83 (76.67)	71 (76.67)	230
4 and more	18 (16.33)	10 (16.33)	21 (16.33)	49
Total	100	100	100	300

Step 1

H_0: The three areas do not differ with regard to the number of bedrooms per residence.

H_1: The three areas differ with regard to the number of bedrooms per residence.

Step 2

Choose $\alpha = 0.05$. Since, in reality, H_0 indicates that factors A and B in Table 15.16 are independent, it follows that under H_0

$$C = \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

is approximately distributed as $\chi^2(4)$.

Step 3

Reject H_0 if $c \geq \chi^2_{4,\,0.05} = 9.488$ or if the p-value ≤ 0.05.

Step 4

When using the observed and expected frequencies given in Table 15.16 it follows that

$$c = \frac{(6-7)^2}{7} + \frac{(7-7)^2}{7} + \ldots + \frac{(21-16.33)^2}{16.33}$$
$$\approx 5.194$$

Step 5

Since $c < 9.488$, H_0 cannot be rejected at a 5% level of significance. The p-value = 0.26796. As for the number of bedrooms per residence, the three random samples can be seen to be homogenous. Thus the three areas do not differ with regard to the number of bedrooms per residence.

Note

If this test is applied to more than two independent samples, and each entity is classified according to either the presence or the absence of a **single** characteristic ($I = 2$), this procedure can be regarded as a test for

$$H_0: \pi_1 = \pi_2 = \ldots = \pi_J,$$

in which $\pi_1, \pi_2, \ldots, \pi_J$ are the parameters of J independent binomial distributions. The $(2 \times J)$ homogeneity test for J independent samples is therefore an extension of the test for the difference between two proportions discussed in Chapter 12, Section 6.4.

5 The log linear analysis of an $(I \times J)$ table

There is a measure of correspondence between the representations of an observed $(I \times J)$ contingency table and an $(r \times k)$ factorial experiment. However, the cell values in an $(I \times J)$ contingency table are observed **frequencies** of the factor categories (A_1, B_1), (A_1, B_2), ..., while the cell values in an analysis of variance experiment such as the $(r \times k)$ factorial experiment are realizations of different treatment combinations applied to one or more experimental units. The analysis of an $(r \times k)$ factorial experiment involves, for the most part, testing the significance of the A and B factors. If the observed averages of the levels of factor A differ significantly, factor A is considered to be significant. The same applies to factor B. In a factorial experiment of this nature, interaction between the A and B factors is also examined, although this is not always of primary importance. On the basis of the well-known factorial experiments in analysis of variance, there have been attempts over the past two decades to develop similar analysis techniques for contingency tables. The interdependence between factors A and B in an $(I \times J)$ contingency table is of primary importance, while the identification of significant differences between the categories of any of the factors is less important.

In 1970 Goodman proposed an analysis technique known as the **log linear analysis technique**, in which the interaction structure between the factors of an

observed two-dimensional, three-dimensional or multi-dimensional contingency table can be analysed in detail. The following exposition explains this analysis technique for a two-way table.

If a random sample of n entities is taken from a population and classified according to two factors, we obtain an observed $(I \times J)$ contingency table as represented in Table 15.13. A similar 'expected' $(I \times J)$ contingency table can be drawn up by replacing the observed numbers o_{ij} by the expected numbers

$$v_{ij} = n \times P(\text{randomly selected entity belongs in cell } (A_i, B_j));$$
$$i = 1, 2, \ldots, I \text{ and } j = 1, 2, \ldots, J.$$

The log linear model representation is now based on

$$\zeta_{ij} = \ell n(v_{ij});\ i = 1, 2, \ldots, I;\ j = 1, 2, \ldots, J$$

in which 'ℓn' indicates the natural logarithm log_e. These unknown quantities, together with their row and column totals, are represented in Table 15.17.

Table 15.17 Logarithms of expected frequencies of an $(I \times J)$ contingency table. The estimates of these unknown values are shown in brackets

	B(column factor)				
A(row fractor)	B_1	B_2	...	B_j	Average
A_1	ζ_{11} (z_{11})	ζ_{12} (z_{12})	... ...	ζ_{1J} (z_{1J})	$\overline{\zeta}_{r1}$ $(\overline{z}_{r1})$
A_2	ζ_{21} (z_{21})	ζ_{22} (z_{22})	... ...	ζ_{2J} (z_{2J})	$\overline{\zeta}_{r2}$ $(\overline{z}_{r2})$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
A_I	ζ_{I1} (z_{I1})	ζ_{I2} (z_{I2})	... ...	ζ_{IJ} (z_{IJ})	$\overline{\zeta}_{rI}$ $(\overline{z}_{rI})$
Average	$\overline{\zeta}_{k1}$ $(\overline{z}_{k1})$	$\overline{\zeta}_{k2}$ $(\overline{z}_{k2})$	... ...	$\overline{\zeta}_{kJ}$ $(\overline{z}_{kJ})$	$\overline{\zeta}$ $\overline{z}$

If $\overline{\zeta}_{ri}$, $\overline{\zeta}_{kj}$ and $\overline{\zeta}$ respectively indicate the row, column and total means of the ζ values, it follows that ζ_{ij} can be expressed as follows:

$$\zeta_{ij} = \overline{\zeta} + (\overline{\zeta}_{ri} - \overline{\zeta}) + (\overline{\zeta}_{kj} - \overline{\zeta}) + (\zeta_{ij} - \overline{\zeta}_{ri} - \overline{\zeta}_{kj} + \overline{\zeta})$$
$$= \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB},\ i = 1, 2, \ldots, I \text{ and } j = 1, 2, \ldots, J$$

Here

$\mu = \bar{\zeta}$, the mean effect

$\lambda_i^A = (\bar{\zeta}_{ri} - \bar{\zeta})$, the effect of A_i

$\lambda_j^B = (\bar{\zeta}_{kj} - \bar{\zeta})$, the effect of B_j

$\lambda_{ij}^{AB} = (\zeta_{ij} - \bar{\zeta}_{ri} - \bar{\zeta}_{kj} + \bar{\zeta})$, the interaction effect between A_i and B_j.

The above linear representation of the logarithms of the expected frequencies $v_{11}, v_{12}, \ldots, v_{IJ}$ is known as the **two-way log linear model**. The effects λ_i^A, $i = 1, 2, \ldots, I$ and λ_j^B, $j = 1, 2, \ldots, J$ are known as the **main effects**, while λ_{ij}^{AB}, $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$ are known as the (first order) **interaction effects**. The unknown μ, together with the λ's are jointly referred to as the **parameters** of the model. From the above model representation it follows, from simple operations that

$$\sum_{i=1}^{I} \lambda_i^A = \sum_{j=1}^{J} \lambda_j^B = 0$$

$$\sum_{i=1}^{I} \lambda_{ij}^{AB} = 0, j = 1, 2, \ldots, J$$

and $$\sum_{j=1}^{J} \lambda_{ij}^{AB} = 0, i = 1, 2, \ldots, I$$

Due to these linear relationships among the main effects and among the interaction effects, it follows that only $(I - 1)$ A effects, $(J - 1)$ B effects and $(I - 1)(J - 1)$ interaction effects are unknown. The model therefore contains

$$1 + (I - 1) + (J - 1) + (I - 1)(J - 1) = IJ$$
$$(\mu) \quad (\lambda^A\text{s}) \quad (\lambda^B\text{s}) \quad (\lambda^{AB}\text{s})$$

unknown parameters. Since there are just as many unknown parameters in the model as cells in the relevant contingency table, the model is known as the saturated model.

The log linear representation of the **saturated model** for an $(I \times J)$ contingency table is given by

$$\ell n(v_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}, \; i = 1, 2, \ldots, I, \; j = 1, 2, \ldots, J$$

in which $v_{11}, v_{12}, \ldots, v_{ij}$ are the expected frequencies, while μ and the λ's represent the parameters of the model.

If there is no interaction between factors A and B, it means that all the λ_{ij}^{AB} effects are equal to zero. In this case the model reduces to the so-called independence model.

The log linear representation of the **independence model** for an $(I \times J)$ contingency table is given by

$$\ell n(v_{ij}) = \mu + \lambda_i^A + \lambda_j^B, \; i = 1, 2, \ldots, I, \; j = 1, 2, \ldots, J$$

For an observed $(I \times J)$ contingency table the log linear analysis technique now involves the following steps:

(a) Decide which of the above models (saturated or independence model) best suits the observed table. This step amounts to testing

H_0: $\lambda_{ij}^{AB} = 0$ for all i and j (abbreviated $\lambda^{AB} = 0$) against
H_1: $\lambda_{ij}^{AB} \neq 0$ for at least one combination (i, j) (abbreviated $\lambda^{AB} \neq 0$).

(b) Estimate all the parameters for the appropriate model. The standardized values of all the estimates give an indication of the importance of the different effects and the significance of these effects are then examined further by means of hypothesis testing.

We shall now discuss these steps separately.

5.1 Decision between saturated and independence models

The test procedure for choosing between these two models is explained in Table 15.18. Many authors prefer the test statistic L to the test statistic C, which is used in the $(I \times J)$ independence test and for a large n has the same distribution as L under H_0. The reason for this is that, as a test statistic, L has certain useful properties that are important in further analyses but will not be discussed here. Note that $E_{11}, E_{12}, \ldots, E_{IJ}$ have the same meaning as the E values used in the test statistic, C, namely estimators for the expected frequencies $e_{11}, e_{12} \ldots, e_{IJ}$ under H_0.

Table 15.18 Testing of H_0: $\lambda^{AB} = 0$ for an observed $(I \times J)$ contingency table

Test statistic	**Distribution of test statistic under H_0**	**Table with critical values**
$L = 2\sum_{i=1}^{I}\sum_{j=1}^{J} O_{ij} \ln\left(\frac{O_{ij}}{E_{ij}}\right)$ with $E_{ij} = \frac{R_i K_j}{n}$	approximately $\chi^2\{(I-1)(J-1)\}$ for n large	Table B.4
Hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
H_0: $\lambda^{AB} = 0$; independence model H_1: $\lambda^{AB} \neq 0$; saturated model	$\ell^* > \chi^2_{(I-1)(J-1),\,\alpha}$	$p = P(L \geq \ell^*)$

(ℓ^* is used as the observed value of L to avoid confusion at a later stage.)

5.2 Estimation of the parameters for the appropriate model

If testing indicates that the saturated model is the best model, $v_{11}, v_{12} \ldots, v_{IJ}$ are estimated by means of the observed $o_{11}, o_{12}, \ldots, o_{IJ}$. On the other hand, if the independence model is a better model for representing an $(I \times J)$ contingency table, the logical estimates for

$v_{11}, v_{12}, \ldots, v_{IJ}$ are the computed $e_{11}, e_{12}, \ldots, e_{IJ}$. Let $\ell_1^A, \ell_2^A, \ldots, \ell_I^A$ and $\ell_1^B, \ell_2^B, \ldots, \ell_J^B$ now designate the estimates of $\lambda_1^A, \lambda_2^A, \ldots, \lambda_I^A$ and $\lambda_1^B, \lambda_2^B, \ldots, \lambda_J^B$ respectively, while $\ell_{11}^{AB}, \ell_{12}^{AB}, \ldots, \ell_{IJ}^{AB}$ are estimates of $\lambda_{11}^{AB}, \lambda_{12}^{AB}, \ldots, \lambda_{IJ}^{AB}$, all based on an observed $(I \times J)$ table.

5.2.1 Saturated model

If the z values in Table 15.17 for all i and j are given by

$$z_{ij} = \ell n(o_{ij})$$

it follows from the relationships between the λ's and the ζ's in the log linear representation of the saturated model that

$$\ell_i^A = (\bar{z}_{ri} - \bar{z}),\ i = 1, 2, \ldots, I$$
$$\ell_j^B = (\bar{z}_{kj} - \bar{z}),\ j = 1, 2, \ldots, J$$
$$\ell_{ij}^{AB} = (z_{ij} - \bar{z}_{ri} - \bar{z}_{kj} + \bar{z}),\quad i = 1, 2, \ldots, I;\ \ j = 1, 2, \ldots, J$$

5.2.2 Independence model

If the z values in Table 15.17 for all i and j are given by

$$z_{ij} = \ell n(e_{ij})$$

it follows from the relations between the λ's and the ζ's in the log linear representation of this model that

$$\ell_i^A = (\bar{z}_{ri} - \bar{z}),\ i = 1, 2, \ldots, I$$
$$\ell_j^B = (\bar{z}_{kj} - \bar{z}),\ j = 1, 2, \ldots, J$$

An estimate of μ is usually not important but can be calculated in any of these models as $\bar{z}$. As was the case for the λ's, we can show that

$$\sum_{i=1}^{I} \ell_i^A = \sum_{j=1}^{J} \ell_i^B = 0$$

$$\sum_{i=1}^{I} \ell_{ij}^{AB} = 0,\ j = 1, 2, \ldots, J$$

$$\sum_{j=1}^{J} \ell_{ij}^{AB} = 0,\ i = 1, 2, \ldots, I$$

To test whether or not any of the estimated parameters represent a significant effect, we use the important Goodman result.

If ℓ is an estimate of λ, one of the unknown parameters in the log linear representation of an $(I \times J)$ contingency table, then

$$\frac{\ell - \lambda}{s_\ell}$$

can for large values of n be regarded as an observation from a n(0, 1) population. Here

$$s_\ell^2 = \begin{cases} \dfrac{1}{I^2 J^2} \displaystyle\sum_{r=1}^{I} \sum_{s=1}^{J} \left(\frac{a_{rs}^2}{o_{rs}} \right) & \text{for the } \textbf{saturated model} \\ \dfrac{1}{I^2 J^2} \displaystyle\sum_{r=1}^{I} \sum_{s=1}^{J} \left(\frac{a_{rs}^2}{e_{rs}} \right) & \text{for the } \textbf{independence model} \end{cases}$$

with the a values given in Table B.13.

The value s_ℓ is merely an estimate of the standard deviation of the estimator of λ so that $\frac{\ell - \lambda}{s_\ell}$ is actually the standard value of the estimate ℓ. If we want to test

H_0: $\lambda = 0$ against H_1: $\lambda \neq 0$,

it is clear that, under H_0 for large n, $\frac{\ell}{s_\ell}$ can be regarded as an observation from a n(0, 1) population. To determine whether or not an effect is significant, the following rule is used:

ℓ is not significant ($\lambda = 0$) if $\left| \frac{\ell}{s_\ell} \right| < 2.58$

ℓ is significant ($\lambda \neq 0$) if $\left| \frac{\ell}{s_\ell} \right| \geq 2.58$

This procedure implies that for each main effect and interaction effect (for the saturated model) the null hypothesis

H_0: $\lambda = 0$ is tested against

H_1: $\lambda \neq 0$

at a 1% level of significance (see Table B.2). Since there are so many estimations involved in this procedure, an effect is only indicated as being significant if the **1% critical value** is exceeded.

The log linear analysis will now be illustrated in the following two examples.

EXAMPLE 15.7

Consider the observed (2×2) table set out in Table 15.10. Table 15.19 now gives the observed o values and also the calculated e and z values. The e values are given in round brackets and the z values in square brackets (in this case $z_{ij} = \ell n(e_{ij})$ – see later).

Table 15.19 Table containing the information required for a log linear analysis of the data given in Table 10.10

A: Property valuation	B: Age		Total
	B_1: Built before 1980	B_2: Built 1980 and later	
A_1: R150 000 or less	22 (16.6822) [2.8143]	29 (34.3178) [3.5357]	51 [6.3500]
A_2: More than R150 000	13 (18.3178) [2.9079]	43 (37.6822) [3.6292]	56 [6.5371]
Total	35 [5.7222]	72 [7.1649]	107 [12.8871]

The decision between the saturated model and the independence model is taken as follows:

Step 1

H_0: $\lambda^{AB} = 0$ (independence model)
H_1: $\lambda^{AB} \neq 0$ (saturated model)

Step 2

Choose $\alpha = 0.01$. If H_0 is true, then

$$L = 2 \sum_{i=1}^{2} \sum_{j=1}^{2} O_{ij} \, \ell n \left(\frac{O_{ij}}{E_{ij}} \right)$$

is approximately distributed according to the $\chi^2(1)$ distribution.

Step 3

Reject H_0 if $\ell^* \geq \chi^2_{1,0.01} = 6.635$ or if the p-value ≤ 0.01.

Step 4

From Table 15.18 it follows that

$$\ell^* = 2\left\{22 \, \ell n\left(\frac{22}{16.6822}\right) + 29 \, \ell n\left(\frac{29}{34.3178}\right) + 13 \, \ell n\left(\frac{13}{18.3178}\right) + 43 \, \ell n\left(\frac{43}{37.6822}\right)\right\}$$

$$= 2\{6.0874 - 4.8827 - 4.4580 + 5.6765\}$$

$$= 4.8465$$

Step 5

Since $\ell^* < 6.635$, H_0 cannot be rejected at a 1% level of significance, in other words, the independence model is selected. The p-value is 0.02770. Note, however, that had we set $\alpha = 0.05$, we would have selected the saturated model.

The model that therefore gives a meaningful representation of the observed (2×2) contingency table based on the above decision is

$$\ell n\,(v_{ij}) = \mu + \lambda_i^A + \lambda_j^B;\ i = 1,\ 2,\ \text{and}\ j = 1,\ 2$$

Let $z_{ij} = \ell n\,(e_{ij})$. It follows from Table 15.19 that

$$\bar{z}_{r1} = \frac{6.3500}{2} = 3.1750$$

$$\bar{z}_{r2} = \frac{6.5371}{2} = 3.2686$$

$$\bar{z}_{k1} = \frac{5.7222}{2} = 2.8611$$

$$\bar{z}_{k2} = \frac{7.1649}{2} = 3.5825$$

$$\bar{z} = \frac{12.8871}{4} = 3.2218$$

From this it follows that

$$\ell_1^A = \bar{z}_{r1} - \bar{z} = 3.1750 - 3.2218 = -0.0468 \quad \text{and} \quad \ell_2^A = \bar{z}_{r2} - \bar{z} = 3.2686 - 3.2218 = 0.0468$$

$$\ell_1^B = \bar{z}_{k1} - \bar{z} = 2.8611 - 3.2218 = -0.3607 \quad \text{and} \quad \ell_2^B = \bar{z}_{k2} - \bar{z} = 3.5825 - 3.2218 = 0.3607$$

Note that $\ell_1^A + \ell_2^A = 0$ and $\ell_1^B + \ell_2^B = 0$, which verifies the calculations. To test which of the estimated effects in this (2×2) table are significant, the associated a values of each estimated effect may be found in Table B.13. Table 15.20 shows the ℓ values, s_ℓ values (independence model) and the standardized estimates. We illustrate the calculation of the s_ℓ values by calculating ℓ_1^A:

$$s_\ell^2 = \frac{1}{2^2 \times 2^2}\left\{\frac{a_{11}^2}{e_{11}} + \frac{a_{12}^2}{e_{12}} + \frac{a_{21}^2}{e_{21}} + \frac{a_{22}^2}{e_{22}}\right\}$$

$$= \frac{1}{16}\left\{\frac{(1)^2}{16.6822} + \frac{(1)^2}{34.3178} + \frac{(-1)^2}{18.3178} + \frac{(-1)^2}{37.6822}\right\}$$

$$\approx 0.0106$$

so that $s_\ell \approx 0.1031$

Table 15.20 Estimated effects, standard deviations of estimates as well as standardized values for the fitting of an independence model to the data in Table 15.10

Effect	Estimate (ℓ)	s_ℓ	$\frac{\ell}{s_\ell}$	Significant?
λ_1^A	−0.0468	0.1031	−0.454	No
λ_2^A	0.0468	0.1031	0.454	No
λ_1^B	−0.3607	0.1031	−3.499	Yes
λ_2^B	0.3607	0.1031	3.499	Yes

In the case of a (2×2) table, the s_ℓ values are the same for all the estimated effects due to the fact that $a_{rs}^2 = 1$ for all values of r and s.

According to Table 15.20 only the column effects are significant. This indicates that the row totals do not differ from one another, but that the column totals do. Significantly more 'new' houses (built 1980 and later) than 'old' houses (built before 1980) were therefore sold.

EXAMPLE 15.8

In Example 15.5 the $(I \times J)$ independence test was used. The factors 'age' and 'number of bedrooms per residence' were found to be strongly interdependent. It is important to determine which of the factor combinations (cells) are chiefly responsible for this dependence. This problem can be best solved by using the log linear analysis. Table 15.21 gives the observed frequencies and also the e and z values used in the log linear analysis. The e values are once again shown in round brackets and the z values in square brackets. The choice between the saturated and independence models is as follows:

Table 15.21 Table containing the information required to carry out a log linear analysis of the data given in Table 15.3

A: Number of bedrooms	B: Age of house					Total
	B_1: Built before 1960	B_2: Built 1960–1969	B_3: Built 1970–1979	B_4: Built 1980–1989	B_5: Built 1990 and later	
A_1: Two bedrooms	41 (17.49) [3.714]	11 (8.01) [2.398]	5 (10.95) [1.609]	9 (18.47) [2.197]	7 (18.07) [1.946]	73 [11.864]
A_2: Three bedrooms	135 (151.69) [4.905]	70 (69.47) [4.249]	92 (94.99) [4.522]	169 (160.20) [5.130]	167 (156.66) [5.118]	633 [23.924]
A_3: Four or more bedrooms	38 (44.81) [3.638]	17 (20.52) [2.833]	37 (28.06) [3.611]	48 (47.33) [3.871]	47 (46.28) [3.850]	187 [17.803]
Total	214 [12.257]	98 [9.480]	134 [9.742]	226 [11.198]	221 [10.914]	893 [53.591]

Step 1

H_0: $\lambda^{AB} = 0$ (independence model)
H_1: $\lambda^{AB} \neq 0$ (saturated model)

Step 2

Choose $\alpha = 0.01$. If H_0 is true, then

$$L = 2 \sum_{i=1}^{3} \sum_{j=1}^{5} O_{ij} \, \ell n \left(\frac{O_{ij}}{E_{ij}} \right)$$

is approximately distributed as $\chi^2(8)$.

Step 3

Since $\chi^2_{8,\,0.01} = 20.090$, it follows that H_0 is rejected if $\ell^* \geq 20.090$ or if the p-values ≤ 0.01.

Step 4

From Table 15.21 it follows that

$$\ell^* = 2\left\{ 41 \ell n \left(\frac{41}{17.49} \right) + 11 \ell n \left(\frac{11}{8.01} \right) + \ldots + 47 \ell n \left(\frac{47}{46.28} \right) \right\}$$

$$= 50.51$$

Step 5

Since $\ell^* \geq 20.090$, H_0 is rejected at a 1% level of significance, and the saturated model will be used to represent this observed (3×5) contingency table. The p-value = 0.000000032.

We now estimate the parameters of the saturated model, namely the parameters of

$$\ell n(v_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}; \quad i = 1, 2, 3; \quad j = 1, 2, \ldots, 5$$

Let $z_{ij} = \ell n \, (o_{ij})$, then from Table 15.21 it follows that

$$\bar{z} = \frac{53.590}{15} = 3.573.$$

The row and column means of the z-values are:

$$\bar{z}_{r1} = 2.373; \; \bar{z}_{r2} = 4.785; \; \bar{z}_{r3} = 3.561; \; \bar{z}_{k1} = 4.086;$$
$$\bar{z}_{k2} = 3.160; \; \bar{z}_{k3} = 3.247; \; \bar{z}_{k4} = 3.733; \; \bar{z}_{k5} = 3.638.$$

The estimated ℓ values are obtained by using the following relationships

$$\ell_i^A = \bar{z}_{ri} - \bar{z}; \quad i = 1, 2, 3$$
$$\ell_j^B = \bar{z}_{kj} - \bar{z}; \quad j = 1, 2, \ldots, 5$$
$$\ell_{ij}^{AB} = z_{ij} - \bar{z}_{ri} - \bar{z}_{kj} + \bar{z}; \; i = 1, 2, 3; \; j = 1, 2, \ldots, 5.$$

The estimates are reflected in the second column of Table 15.22. The a values for a (3×5) table, as reflected in Table B.13, are used to calculate the standardized values. These a values are then used to calculate the standard deviations ℓ for each estimate. The following example of ℓ_{12}^{AB} illustrates these calculations.

$$s_\ell^2 = \frac{1}{3^2 \times 5^2}\left\{\begin{array}{l} \dfrac{(-2)^2}{41} + \dfrac{(8)^2}{11} + \dfrac{(-2)^2}{5} + \dfrac{(-2)^2}{9} + \dfrac{(-2)^2}{7} \\ + \dfrac{(1)^2}{135} + \dfrac{(-4)^2}{70} + \dfrac{(1)^2}{92} + \dfrac{(1)^2}{169} + \dfrac{(1)^2}{167} \\ + \dfrac{(1)^2}{38} + \dfrac{(-4)^2}{17} + \dfrac{(1)^2}{37} + \dfrac{(1)^2}{48} + \dfrac{(1)^2}{47} \end{array}\right\}$$
$$\approx 0.04012,$$

so that $s_\ell \approx 0.20030$

In addition to the estimated effects, Table 15.22 also shows the s_ℓ values and the standardized estimates. Although we are not primarily interested in the main effects, we note that four of the eight are in fact significant at a 1% level. This was to be expected, since the row totals and the column totals differ substantially. Of interest here, however, are the large positive and negative standardized ℓ values, namely ℓ_{11}^{AB}, ℓ_{21}^{AB} and ℓ_{31}^{AB}. We therefore conclude that cells (A_1, B_1), (A_2, B_1) and (A_3, B_1) are responsible for the significant interaction between factors A and B. In contrast to the negative signs of ℓ_{21}^{AB} and ℓ_{31}^{AB}, the positive sign of ℓ_{11}^{AB} indicates that significantly more two-bedroomed residences were built in the years prior to 1960.

Table 15.22 Estimated λ effects, standard deviations of ℓ estimates and standardized ℓ values for the log linear analysis of the saturated model for the data in Table 15.3

Effect	ℓ	s_ℓ	$\frac{\ell}{s_\ell}$	Significant?
λ_1^A	−1.200	0.10490	−11.44	Yes
λ_2^A	1.212	0.06330	19.15	Yes
λ_3^A	−0.012	0.07396	−0.16	No
λ_1^B	0.513	0.08537	6.01	Yes
λ_2^B	−0.413	0.11978	−3.45	Yes
λ_3^B	−0.326	0.13883	−2.35	No
λ_4^B	0.160	0.11227	1.43	No
λ_5^B	0.065	0.12147	0.54	No

Tabel 15.22 (continued)

Effect	ℓ	s_ℓ	$\frac{\ell}{s_\ell}$	Significant?
λ_{11}^{AB}	0.828	0.14056	5.89	Yes
λ_{12}^{AB}	0.438	0.20030	2.19	No
λ_{13}^{AB}	−0.438	0.25858	−1.69	No
λ_{14}^{AB}	−0.336	0.20495	−1.63	No
λ_{15}^{AB}	−0.492	0.22565	−2.18	No
λ_{21}^{AB}	−0.393	0.09676	−4.06	Yes
λ_{22}^{AB}	−0.123	0.13341	−0.92	No
λ_{23}^{AB}	0.063	0.14846	0.42	No
λ_{24}^{AB}	0.185	0.11992	1.54	No
λ_{25}^{AB}	0.268	0.12863	2.08	No
λ_{31}^{AB}	−0.436	0.12086	−3.61	Yes
λ_{32}^{AB}	−0.315	0.16784	−1.88	No
λ_{33}^{AB}	0.376	0.16351	2.30	No
λ_{34}^{AB}	0.150	0.13721	1.09	No
λ_{35}^{AB}	0.224	0.14515	1.54	No

Notes

(a) We have discussed only the two most important log linear model representations for the $(I \times J)$ contingency table. Where the independence model is preferred to the saturated model, this does not necessarily mean that this model is the simplest representation of an observed $(I \times J)$ table. Strictly speaking, any one of the following three models may provide a simpler representation for such an observed model:

(i) $\ell n\,(v_{ij}) = \mu + \lambda_i^A$

(ii) $\ell n\,(v_{ij}) = \mu + \lambda_j^B$

(iii) $\ell n\,(v_{ij}) = \mu$

Model (i) implies that there is no difference in the incidence of the B categories in the population. On the other hand, model (ii) implies that there is no difference in the incidence of the A categories in the population. The simplest model, namely (iii), means that there is an equal probability that an entity can be placed in any one of the $(I \times J)$ cells of the contingency table. The standardized values of the estimated effects in the independence model give some indication as to which of the above three models, if any, is applicable. The analysis of Example 15.7 shows that, if the independence model is preferred to the saturated model, model (ii) will probably provide an even better representation of the observed (2×2) contingency table.

This deduction follows from the fact that the A effects in Example 15.7 do not deviate significantly from zero. If the independence model is accepted, the significance of any one of the above models can be tested and the required estimates made as explained above.

(b) If one or more of the observed frequencies is zero, problems arise in a log linear analysis. The reason for this is that $\ell n\,(0) = -\infty$ so that the test statistic L is not defined. Various writers recommend that, in such a case, a small positive number (say 0.5) should be added to each observed frequency. Alternatively, adjacent rows and/or columns can be combined to eliminate zero frequencies (see Section 4.1, Note (a)). Obviously, the best solution is to eliminate the zeros by drawing a larger sample.

(c) Since the main effects are usually less important than the interaction effects, the parameters in a log linear analysis are often estimated only when the saturated model is used as a representation.

6 In conclusion

6.1 Multi-dimensional contingency table

If the interaction structure of a three-way or multi-dimensional contingency table needs to be analysed, the log linear analysis technique can still be used. In such cases – apart from first order effects – often also second and higher order interaction effects are present. As a result, the selection of a model and the estimation of the parameters becomes more complicated. Despite these complexities, the log linear analysis technique remains basically unchanged. The log linear analysis technique was especially designed to analyse multivariate data sets that are partly or totally qualitative.

6.2 Computer programs

Standard computer programs in the BMDP, SAS and SPSS series can be used if the analyses of an $(I \times J)$ contingency table become unwieldy. The BMDP and SAS series offer a useful program for the log linear analysis of multi-dimensional contingency tables. Log linear analyses are so comprehensive that it is virtually impossible to compute them with a standard pocket calculator.

Exercises

1 The following table was obtained by drawing up a (1×10) contingency table for the random numbers in Table B.1, second page rows 26 and 27:

Distribution of random numbers										
0	1	2	3	4	5	6	7	8	9	Total
7	11	7	12	14	12	6	9	8	14	100

Test at a 5% level of significance whether the random numbers do appear in Table B.1 with an equal probability.

2 The frequency distribution below reflects test marks (out of 40) scored by students in a statistics test. Could these marks possibly have been taken from a normal population? Assume $\mu = 28$ and $\sigma^2 = 7.5$ and test at a 5% level of significance.

Statistics test marks	
Class intervals	**Number**
[6, 11)	5
[11, 16)	6
[16, 21)	12
[21, 26)	19
[26, 31)	21
[31, 36)	9
[36, 40)	8

3 The number of lorries per hour loading bricks at a brickworks was observed for 200 working hours. The table below reflects a (1×10) contingency table of the number of hours during which 0, 1, 2, etc. lorries arrived. Test at a 5% level of significance whether X = number of lorries arriving to load bricks each hour might have a Poisson distribution.

Number of lorries x	0	1	2	3	4	5	6	7	8	9–15	**Total**
Number of 1 hour periods	1	7	16	30	33	37	30	18	12	16	200

4 In two consecutive weeks 120 patients with the same disease are subjected to two medical treatments. The (2×2) observed contingency table below shows the reactions to the treatments:

	Treatment B	
Treatment A	Favourable	Unfavourable
Favourable	23	36
Unfavourable	32	29

(a) Is there dependence (interaction) between the two treatments? $(\alpha = 0.01)$

(b) Test whether the degree of success for the two treatments differs $(\alpha = 0.01)$.

5 The first 250 residences in Johannesburg (Appendix A, Data Set 1) are classified in a (3×3) contingency table according to selling price (R1 000) and size (m^2).

Selling price (A) (R1 000)	Size in square metres (B) [80, 150)	[150, 200)	[200, 350)
[25, 150)	59	39	29
[150, 200)	10	21	22
[200, 501)	4	14	52

Test at a 5% level of significance whether factors A and B are independent.

6 After a general election 150 voters are approached about their support or lack thereof for Party A in this election and also in the previous election. Only 119 of the voters approached were prepared to provide this information. The (2×2) contingency table below is a summary of the information:

Present election	Previous election For Party A	Against Party A
For Party A	42	28
Against Party A	28	21

(a) Is there a dependence between the voting pattern for the two elections $(\alpha = 0.05)$?

(b) Does the support for Party A in the present election differ from the support in the previous election $(\alpha = 0.05)$?

7 There is a suspicion that the academic achievements of house committee members of women's residences are unsatisfactory because they devote too much time to residence duties. From all the women's residences at a university, 96 senior female students were involved in an investigation into the matter.

The table below contains a classification of the academic achievements of the women who served on house committees and those who did not serve on house committees.

Academic achievement	House committee member Yes	No
Good	8	14
Average	14	25
Poor	15	20

Are committee membership and academic achievement independent? (Test at a 5% level of significance.)

8 The following table contains a classification of industrial accidents according to the degree of seriousness and the work shift in which they occurred:

Work shift	Degree of accident Minor	Mild	Serious
Morning	10	6	7
Afternoon	18	14	7
Night	22	17	21

Apply a suitable test to determine whether there is a dependence between the degree of seriousness of an accident and the work shift during which it occurs ($\alpha = 0.05$).

9 An educational researcher wants to determine whether there is a relationship between scholastic achievement and the number of hours that children spend watching television programmes. In a survey of 90 pupils he obtained the following results:

Achievement	Average number of hours in front of TV		
	Less than 2	2 to 7	More than 7
Passed	13	32	15
Failed	4	8	18

What conclusion can you draw ($\alpha = 0.05$)?

10 Analyse the contingency table in Exercise 6 with the aid of the log linear analysis technique.

11 Use the log linear analysis technique to analyse the contingency tables for the data in:
(a) Exercise 5;
(b) Exercise 7;
(c) Exercise 8; and
(d) Exercise 9.

12 A random sample of eighty residences in Durban (Appendix A, Data Set 1) has the following building areas (m^2):

Building areas of Durban residences	
Class intervals	Number
[70, 110)	10
[110, 150)	14
[150, 190)	18
[190, 230)	16
[230, 270)	12
[270, 310)	6
[310, 350)	4
	80

Test at a 1% level of significance whether these areas could possibly be regarded as coming from a normal population.

13 A stratified random sample of families is drawn from the most popular suburbs in a city and the ages of the male heads of families are noted. The table below reflects the information once the ages have been divided into four classes. Test at a 5% level of significance whether the three suburbs

can be regarded as homogeneous with regard to the age of the male heads of families.

Age (years)	Suburbs		
	Panorama	Gardens	Welbekend
Younger than 30	10	34	11
[30, 40)	29	36	23
[40, 50)	18	16	14
50 and older	26	9	30

14 The table below classifies 250 men according to place of work and place of birth. Test whether men who grow up in rural areas prefer to work in the government or in the private sector. Use a 5% level of significance.

Place of birth	Place of work	
	Government sector	Private sector
Rural areas	80	34
City	92	44

15 The table below shows casualties resulting from road traffic accidents on South African roads in 1990 with respect to sex and age. The information was obtained from *SA Statistics 1992*.

Age (years)	Type of casualty			
	Deaths		Injuries	
	Male	Female	Male	Female
0–9	600	364	4 273	2 557
10–19	1 001	315	11 038	4 272
20–29	2 913	682	31 106	9 931
30–39	2 088	477	19 432	6 247
40–49	1 249	308	13 049	4 207
50–59	615	172	5 869	3 518
60 and older	403	170	3 492	1 585

(a) Test at a 5% level of significance whether there is a dependence between gender and age with respect to the number of deaths.
(b) Is there a dependence between the type of casualty and age ($\alpha = 0.01$)?
(c) Does the number of male deaths differ from the number of female deaths ($\alpha = 0.05$)?
(d) Do the age groups differ with respect to the number of deaths ($\alpha = 0.01$)?

16 The table below indicates the duration of marriages (in years) before divorce of whites in the years 1980, 1985 and 1990:

Year of divorce	Duration of marriage (in years)			
	Less than 1 year	1–2 years	3–5 years	More than 5 years
1980	782	2 556	3 369	9 836
1985	907	3 034	3 791	9 303
1990	877	3 426	4 269	11 459

(a) Test for differences in the number of divorces in the three years ($\alpha = 0.05$).

(b) Is there a dependence between the duration of the marriage and the year of divorce? If so, indicate the significant cells by using a 1% level of significance.

17 A personnel officer investigates the leave taken by 64 workers over a period of one year. The data below reflects the number of days' leave and the types of appointment of workers.

Type of appointment	Number of days' leave per year		
	Less than 6 days	6 to 12 days	More than 12 days
Full time	12	9	9
Part time	7	19	8

Test for independence between the number of days' leave and the type of appointment of the workers ($\alpha = 0.05$).

18 The hydrogen sulphide content (volume percentage) of a flow of gas is monitored on an ongoing basis. At the end of the day a mean H_2S concentration is calculated and stored in a database for later use. Research has indicated that it would be economically justified to recycle a certain component in the gas. An important requirement is that the H_2S content must be normally distributed and that there must be 95% certainty that it lies between 1.0% and 2.0% (volume). The table below shows the H_2S concentration for 1 191 days.

Class boundaries	Frequency
[0.7, 0.8)	1
[0.8, 0.9)	5
[0.9, 1.0)	5
[1.0, 1.1)	16
[1.1, 1.2)	62
[1.2, 1.3)	132
[1.3, 1.4)	212
[1.4, 1.5)	243
[1.5, 1.6)	222
[1.6, 1.7)	149
[1.7, 1.8)	75
[1.8, 1.9)	41
[1.9, 2.0)	20
[2.0, 2.1)	6
[2.1, 2.2)	1
[2.2, 2.3)	1
Total	1191

Data obtained from MBL Group SEC00193

Determine from this information whether it would be justified to implement the project ($\alpha = 0.01$)?

19 In a market research questionnaire on the perception of video cassette consumers in South Africa the respondents were asked the following question:

Would you buy a video cassette manufactured in South Africa?

Never	Seldom	Sometimes	Often	Always

The results of the completed questionnaires are given below for men and for women.

Answer to question	**Sex**	
	Male	Female
Never	5	0
Seldom	6	6
Sometimes	43	14
Often	10	15
Always	12	9

Combine the categories 'Never' and 'Seldom' and test the hypothesis that the answer to the question is independent of gender. Use $\alpha = 0.05$.

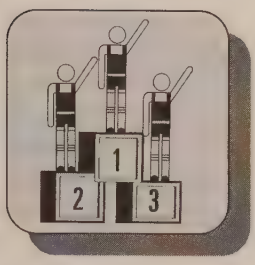
20 You are given the following contingency table. The rows represent the frequency of use of frozen vegetables. The columns indicate the number of children in the families of the respondents (Appendix A, Data Set 2).

Consumption of frozen vegetables	**Number of children**			
	No children	One child	Two children	Three and more
Not regularly	17	31	48	35
Regularly	58	43	105	78

Test the null hypothesis that the consumption of frozen vegetables is independent of the number of children in a family by using $\alpha = 0.05$.

CHAPTER 16

Distribution-free methods



1 Introduction

In the general framework of statistical inference the nature of the observed data often creates problems. Take, for example, ranking of some or other type of sport in which players from different clubs are ranked according to the quality of their game. The question now is whether the players from one club are generally better than those from another, judging from the positions of the players in the ranking. Suppose, for example, the players in Club A occupy positions 2, 4, 5, 6, 8 and 9, and the players from Club B positions 1, 3, 10, 15 and 16, with the rest of the positions taken by players from other clubs. Can we say that the players from the two clubs are on a par with each other? The nature of the observed data is **ordinal** (see Chapter 1, Section 3) and it is, strictly speaking, not justifiable to apply the well-known t test for two independent samples (see Chapter 12, Section 6.2), since the observed values are not taken from normal populations.

Problems such as these have led to the development of the so-called **distribution-free** tests. The term 'distribution-free' refers to the assumptions with regard to the distributions of the underlying populations from which the samples are drawn. The implication is that the application of such procedures does not depend on strict assumptions with respect to the distributions. The most important advantage of distribution-free tests is therefore that they can be used where, strictly speaking, the traditional tests discussed in previous chapters cannot be applied.

A second advantage of distribution-free tests is that they are designed for use in nominal, ordinal and quantitative data. With quantitative data the observations are divided into categories or arranged according to size before the tests are applied. In an investigation into the effectiveness of a painkiller, patients can, for instance, simply indicate whether or not it relieves pain. In a beauty competition with ten participants and five judges, each judge can arrange the ten participants in order of preference so that 'measuring' here means arranging.

A third advantage of distribution-free tests is that the test procedures are usually easy to understand and that the calculations required are simple. Consequently a good theoretical background in statistics is normally not a prerequisite for anyone using these tests.

However, distribution-free tests do not have only advantages. As mentioned above, the tests can also be used for quantitative data once the observations have been divided into categories or arranged according to size. It goes without saying that, as a result of this categorization we do not fully utilize all information contained in quantitative data. In cases where strict assumptions about the underlying distribution do apply, this loss of information means that a distribution-free test does not have the same ability to distinguish between two hypotheses as does the traditional test under these assumptions.

Finally, we must remember that the application of distribution-free tests does require certain assumptions regarding the underlying population distributions. The assumptions concerned will be given with the discussion of each test. In this chapter we shall deal with a wide variety of problems for which distribution-free tests can be applied. In most cases we shall discuss only one such test, although in some instances there are a number of alternative tests to be found in the literature on statistics.

The scheme we shall follow involves one-sample and two-sample problems (see Chapter 12, Sections 5 and 6), the multi-sample problem (see Chapter 14) and the problem of regression and correlation (see Chapter 13). While reading this chapter the reader can refer to these chapters for perspective on the new tests. In conclusion, a test for randomness and a few aspects of estimation within the framework of distribution-free methods will be discussed.

2 The one-sample problem (testing the population median)

As stated in Chapter 4, Section 2, the median is a more representative measure of location than the arithmetic mean in an observed data set with a skewed frequency distribution. The population median, say η, is defined by

$$P(X \leq \eta) = 0.5$$

As in the case of an observed data set, η will be a more representative measure of location than the population mean, μ, if the theoretical distribution of the stochastic variable being studied is skewed. For this reason hypotheses on η often have to be tested, rather than hypotheses on μ. Take, for example, the income distribution of a population of workers. It is common knowledge that the distribution is positively skewed so that the mean income of the population is not a representative value of the income distribution and that η should rather be used.

2.1 The sign test

The sign test is the oldest and best-known test that can be used to test hypotheses pertaining to the population median. Table 16.1 contains an exposition of the procedure for testing

H_0: $\eta = \eta_0$

against the various possible alternatives.

Table 16.1 Testing H_0: $\eta = \eta_0$, on the basis of a random sample $X_1, X_2, \ldots, X_n$ from a population with median η

Test statistic	**Distribution of test statistic under H_0**	**Table with critical values**
B = number of X's greater than η_0	n small: B is distributed bi$(n, \frac{1}{2})$	Table B.10
$Z = \dfrac{B - \frac{n}{2}}{\sqrt{\frac{n}{4}}}$	n large: Z approximately has a n(0, 1) distribution	Table B.2
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
H_1: $\eta \neq \eta_0$	$b \leq b_{n,\alpha/2}$ or $b \geq n - b_{n,\alpha/2}$	$2 \min \{P(B \leq b), P(B \geq b)\}$
H_1: $\eta > \eta_0$	$b \geq n - b_{n,\alpha}$	$P(B \geq b)$
H_1: $\eta < \eta_0$	$b \leq b_{n,\alpha}$	$P(B \leq b)$

The statistic B is known as the **sign test statistic**.

2.1.1 Tables for the sign test

Since the sign test statistic B has a bi$(n, 0.5)$ distribution under H_0, tables of the binomial distribution can be used to find critical values for the sign test. The critical values can, for example, be calculated from Table B.6 for the tabulated values of n at $\pi = 0.5$. For convenience the critical values were calculated beforehand and these critical values $b_{n,\alpha}$ of the sign test for different values of n and α appear in Table B.10. The $b_{n,\alpha}$ is defined as the largest integer for which, under H_0,

$$P(B \leq b_{n,\alpha}) \leq \alpha.$$

It should be noted that the discrete nature of the binomial distribution usually prevents calculation of the b value for which the above probability will be exactly equal to α. Take, for example, the case where $n = 20$ and $\alpha = 0.05$. From Table B.10, $b_{20,0.05} = 5$. However, from Table B.6 it follows that for X, a bi(20, 0.5) stochastic variable,

$$P(X \leq 5) = 0.02069$$
$$\text{and } P(X \leq 6) = 0.05766.$$

This means that $b_{20,0.05}$ is the largest value of b for which $P(B \leq b) \leq 0.05$ under H_0. Figure 16.1 illustrates this aspect. The value $b_{n,\alpha}$ is known as the

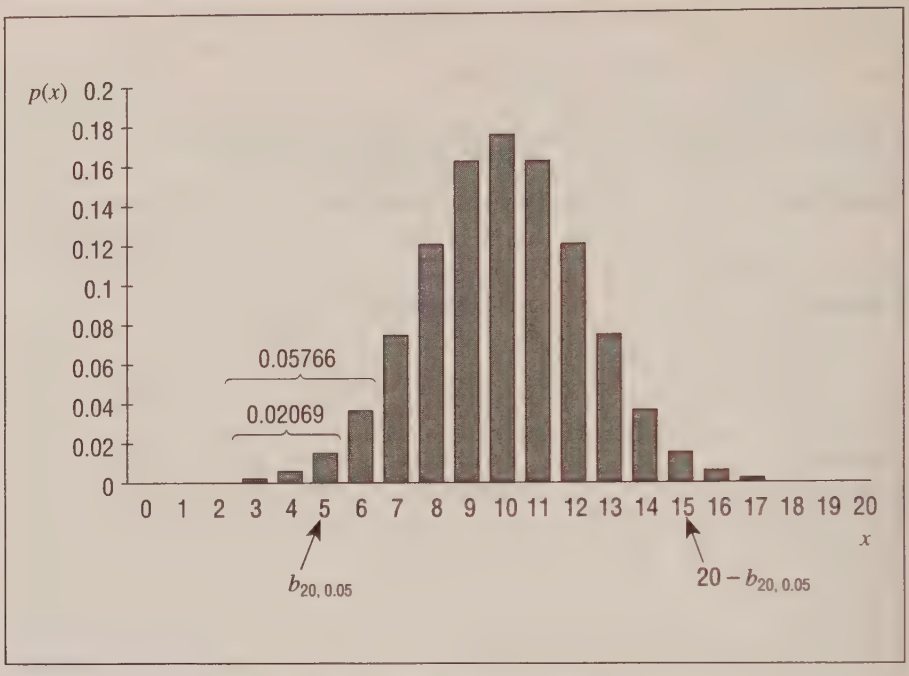


Figure 16.1 Determining the critical values of the sign test

left critical value of the sign test. From the symmetry of the bi(n, 0.5) distribution it follows that the **right critical value** is $n - b_{n,\alpha}$.

Thus it follows that the left and the right critical values for $n = 20$ and $\alpha = 0.05$ are $b_{20,\,0.05} = 5$ and $20 - b_{20,\,0.05} = 15$. Figure 16.1 also illustrates this relation.

The tables for the sign test are given only for $n \leq 40$, since the normal approximation of B under H_0 is already satisfactory for $n > 40$ (see Table 16.1).

EXAMPLE 16.1

A survey of the annual income of 30 randomly selected staff members of a large firm rendered the results in Table 16.2. The arithmetic mean of the 30 values is $\bar{x} = 79.72$ thousand rand and the median is the mean of the 15th and 16th smallest values (see Chapter 4), namely (55.20 + 58.40)/2 = 56.8 thousand rand. This points to a relatively skewed distribution of incomes. The population mean μ can be estimated at approximately R80 000, and the question is whether this estimate is a representative figure for employee income. This would be the case if the population median were also approximately R80 000. This issue is addressed with the aid of the sign test.

Table 16.2 Annual income (R1 000) of 30 persons employed by a large firm

Person	Income	Person	Income	Person	Income
1	163.8	11	200.0	21	32.6
2	3.8	12	165.2	22	59.8
3	84.2	13	58.4	23	12.8
4	172.4	14	106.4	24	202.8
5	336.0	15	10.0	25	78.8
6	65.4	16	211.6	26	55.2
7	3.4	17	12.0	27	38.0
8	15.6	18	48.6	28	36.8
9	61.2	19	29.2	29	4.6
10	41.2	20	78.2	30	3.6

Step 1

H_0: $\eta = 80.0$
H_1: $\eta \neq 80.0$

Step 2

Select $\alpha = 0.05$. The test statistic
B = number of values larger than 80.0
has a bi(30,0.5) distribution under H_0.

Step 3

Reject H_0 if $b \leq b_{30,\,0.025}$ or $b \geq 30 - b_{30,\,0.025}$,
that is, if $b \leq 9$ or $b \geq 21$.
Alternatively: Reject H_0 if the p-value ≤ 0.05.

Step 4

From Table 16.2 it follows that there are 9 observations larger than 80, so that $b = 9$.

Step 5

Since $b \leq 9$, H_0 is rejected at a 5% level of significance. Alternatively, it follows from Table B.6 of Appendix B that the p-value = 2 min $\{P(B \leq 9), P(B \geq 9)\}$ = 0.04278 < 0.05 so that H_0 is rejected at a 5% level of significance. The median income of the staff in the firm therefore apparently differs from R80 000.

EXAMPLE 16.2

In Example 12.9 the board of directors of a particular firm needed to establish whether the mean first-year sales figures of representatives who had completed a new training course exceeded R300 000. In the example we assumed that the first-year sales figures had a normal distribution. From a normal probability plot of the

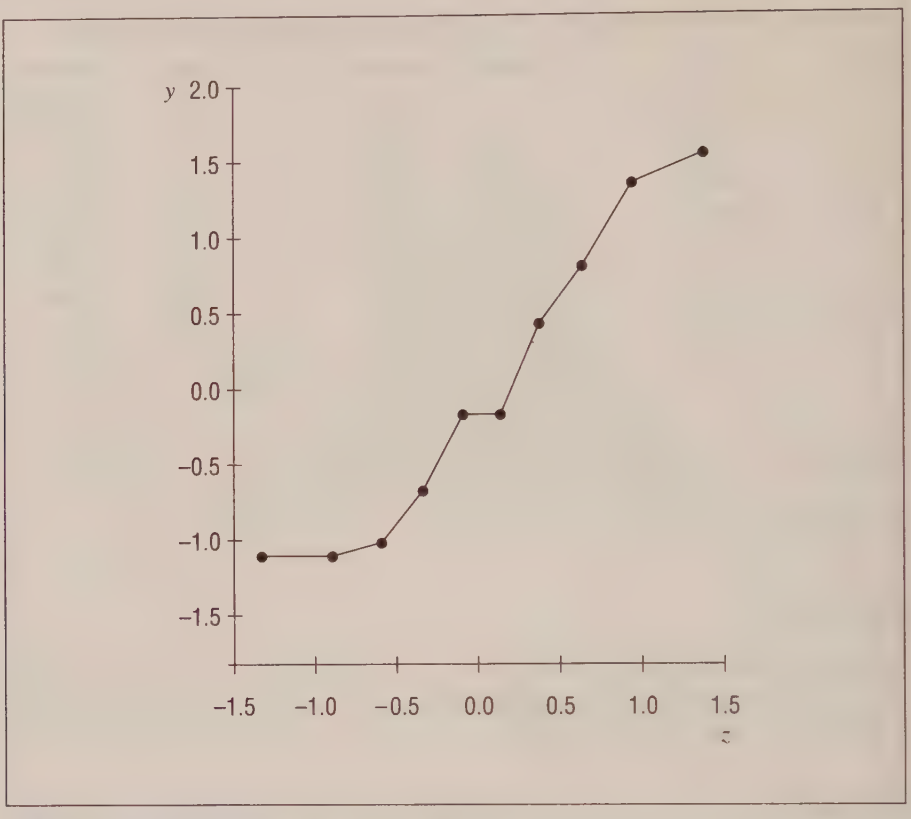


Figure 16.2 Normal probability plot of sales figures

data (see Figure 16.2) it appears that the underlying population is possibly not normally distributed, but relatively symmetrical. However, it is important to know whether the new training course will raise the median of first-year sales significantly higher than R300 000. The 10 observations (in rand) were the following:

287 900	419 400	338 300	287 500	310 850
292 600.	390 050	369 850	430 400	338 450

Step 1

H_0: $\eta = 300\,000$
H_1: $\eta > 300\,000$

Step 2

Again select $\alpha = 0.05$. Under H_0
B = number of sales figures above 300 000
has a bi(10, 0.5) distribution.

Step 3

Reject H_0 if $b \geq 9$ or if the p-value ≤ 0.05.

Step 4

From the observations given in Example 12.9 it follows that 7 of the sales figures are higher than R300 000, so that $b = 7$.

Step 5

Since $b < 9$, H_0 is not rejected at a 5% level of significance. The observed data and use of the sign test would not therefore lead to the conclusion reached in Example 12.9. The p-value $= P(B \geq 7) = 1 - P(B \leq 6) = 1 - 0.82813 = 0.17187 > 0.05$.

Notes

(a) As seen in Chapter 12, the t test is not sensitive to medium deviations from normality by the data. The application of the t test on the first-year sales data is therefore justified. In Example 12.9 H_0: $\mu = 300\,000$ is rejected in favour of H_1: $\mu > 300\,000$. Example 16.2 therefore illustrates very clearly the price that has to be paid for the loss of information if the sign test is used, while the application of the t test is justified. The sign test should consequently never be used out of hand in place of the t test.

(b) The sign test can be expanded to test hypotheses on any population quantile. Thus, for example, the test statistic

B = number of X's greater than a pre-specified value of the third population quartile,

will have a bi(n, 0.25) distribution under H_0.

(c) When one or more of the observed values $x_1, x_2, \ldots, x_n$ are exactly equal to η_0, these values are not taken into account and the sign test is applied to the remaining x values. In such a case B therefore has a bi($n-k$, 0.5) distribution, where k = number of x values equal to η_0.

2.2 The test for symmetry

From the above discussion of the sign test for H_0: $\eta = \eta_0$ it is clear that only the number of X values larger than η_0 determines the value of B. The relative size of the differences $(X_i - \eta_0)$ has no influence. In 1945 F. Wilcoxon proposed a test that takes the relative sizes into account. This test is commonly known as the Wilcoxon test for symmetry, or the signed rank test. Before giving an explanation of this test procedure, we must explain the concept of ranking.

The position of x_i in the ranking of n values $x_1, x_2, \ldots, x_n$ is known as the **rank** of the value and is designated by r_i.

Take, for example, the following observed sample:

x_1	x_2	x_3	x_4	x_5
4.2	3.9	–2.1	4.6	–6.3

The array of the x values is given by:

x_5	x_3	x_2	x_1	x_4
–6.3	–2.1	3.9	4.2	4.6

and the ranks by

r_1	r_2	r_3	r_4	r_5
4	3	2	5	1

In this example it is clear that r_1 is the rank of x_1, r_2 is the rank of x_2, etcetera. When a second sample of x values is observed, it is logical that x_5 will not necessarily again be the smallest, x_3 the second smallest, et cetera. We note, therefore, that $r_1, r_2, \ldots, r_5$ are observed values of the stochastic variables $R_1, R_2, \ldots, R_5$, in which the R values are associated with the random sample $X_1, X_2, \ldots, X_5$. Note that for every sample of size 5

$$\begin{aligned}\sum_{i=1}^{5} r_i &= 4 + 3 + 2 + 5 + 1 \\ &= 1 + 2 + 3 + 4 + 5 \\ &= 15\end{aligned}$$

Generally, for a sample of size n,

$$\begin{aligned}\sum_{i=1}^{n} r_i &= 1 + 2 + \ldots + n \\ &= \frac{n(n+1)}{2}\end{aligned}$$

To apply the signed rank test the absolute differences $|X_1 - \eta_0|$, $|X_2 - \eta_0|$, $\ldots$, $|X_n - \eta_0|$ have to be determined and an array must be drawn up for the absolute values. This test is now based on the sum of the ranks of the absolute differences associated with positive values $(X_i - \eta_0)$, namely

$$\sum_{X_i - \eta_0 > 0} R_i = \sum_{(+)} R_i, \text{ with } R_i \text{ the rank of } |X_i - \eta_0|$$

Table 16.3 explains the test procedure. It is important to note that the signed rank test is based on the assumption that the sample of observations is taken from a symmetrically distributed population. This test therefore tests both

590 H_0: $\eta = \eta_0$ and H_0: $\mu = \mu_0$

Table 16.3 Testing H_0: $\eta = \eta_0$ on the basis of a random sample $X_1, X_2, \ldots, X_n$ from a population that is symmetrically distributed around its median η

Test statistic	Distribution of test statistic under H_0	Table with critical values
$S = \sum_{(+)} R_i$, R_i = rank of $\lvert X_i - \eta_0 \rvert$	n small: S has an exact distribution	Table B.11
$Z = \dfrac{S - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$	n large: Z approximately has a n(0, 1) distribution	Table B.2
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-values**
H_1: $\eta \neq \eta_0$	$s \leq s_{n,\alpha/2}$ or	2 min $\{P(S \leq s),$
	$s \geq n(n+1)/2 - s_{n,\alpha/2}$	$P(S \geq s)\}$
H_1: $\eta > \eta_0$	$s \geq n(n+1)/2 - s_{n,\alpha}$	$P(S \geq s)$
H_1: $\eta < \eta_0$	$s \leq s_{n,\alpha}$	$P(S \leq s)$

2.2.1 Tables for the signed rank test

Under H_0, the test statistic S of the signed rank test has a discrete distribution that is symmetrical around $\mu_s = n(n+1)/4$. There are also tables of critical values for the signed rank test.

In Table B.11, $s_{n,\alpha}$ is tabulated for different values of n and α, in which $s_{n,\alpha}$ is the largest integer for which, under H_0,

$$P(S \leq s_{n,\alpha}) \leq \alpha$$

The value $s_{n,\alpha}$ is the left critical value of the signed rank test. From the symmetry of the distribution of S under H_0 it follows that the right critical value of this test is given by $n(n+1)/2 - s_{n,\alpha}$. Thus, for example, for $n = 20$ and $\alpha = 0.05$ it follows that the left and right critical values are given by

$$s_{20,\,0.05} = 60 \text{ and } \frac{20(20+1)}{2} - s_{20,\,0.05} = 150$$

EXAMPLE 16.3

Turn again to Example 12.9 and consider the ten first-year sales figures of the representatives who were exposed to a new training programme. Suppose we only assume that the first-year sales have a symmetrical distribution. Table 16.4 reflects the ten sales figures and also contains the columns required to calculate the test statistic S to test H_0: $\eta = 300\,000$. In the last column the sign of the difference, $(x_i - 300\,000)$, is indicated in brackets. The test procedure is as follows:

Table 16.4 Calculation table for applying the test for symmetry to the data in Example 12.9

Sales figures x_i	Differences $(x_i - 300\,000)$	Absolute differences $\lvert x_i - 300\,000 \rvert$	Ranks of absolute differences r_i
287 900	–12 100	12 100	3 (–)
419 400	119 400	119 400	9 (+)
338 300	38 300	38 300	5 (+)
287 500	–12 500	12 500	4 (–)
310 850	10 850	10 850	2 (+)
292 600	–7 400	7 400	1 (–)
390 050	90 050	90 050	8 (+)
369 850	69 850	69 850	7 (+)
430 400	130 400	130 400	10 (+)
338 450	38 450	38 450	6 (+)

Step 1

H_0: $\mu = 300\,000$
H_1: $\mu > 300\,000$

The following set of hypotheses is equivalent to this:

H_0: $\eta = 300\,000$
H_1: $\eta > 300\,000$

Step 2

Again select $\alpha = 0.05$. The test statistic is

$$S = \sum_{(+)} R_i$$

Step 3

Reject H_0 if $s \geq \frac{(10 \times 11)}{2} - s_{10,0.05} = 45$ or if the p-value ≤ 0.05.

Step 4

$s = 9 + 5 + 2 + 8 + 7 + 10 + 6$
$= 47$

Step 5

Since $s \geq 45$, H_0 is rejected at a 5% level of significance. We conclude, as in Example 12.9, that the mean (or median) sales figures of the representatives who followed the new course exceeds R300 000. From a table of the cumulative distribution function of S it follows that the p-value $= P(S \geq 47) = 0.0244 < 0.05$.

Notes

(a) The decision to reject H_0, as taken in Example 16.3, corresponds with the decision made in Example 12.9, but differs from the one in Example 16.2. A calculation of the median, the arithmetic mean and the standard deviation of the sales figures concerned gives

$$me = 338.375,\ \bar{x} = 346\,530 \text{ and } s = 53\,768, \text{ so that } \frac{\bar{x} - me}{s} = 0.15$$

This indicates that the observed sample was drawn from a population with a relatively symmetrical distribution. In view of this and also because the signed rank test utilizes more information than the sign test, we prefer the decision obtained by using the signed rank test to the one obtained by means of the sign test. In fact, it can be theoretically proved that, in the case of sampling from a symmetrical distribution, the signed rank test is better able to distinguish between H_0 and H_1 than is the sign test. However, the sign test is valuable when testing hypotheses on the population median in populations with a skewed distribution, since the signed rank test may not be used in such cases. In contrast to the sign test, the signed rank test can also not be used to test hypotheses on other population quantiles (see Note (b) on the sign test).

(b) As in the case of the sign test, the x values for which $(x_i - \eta_0) = 0$ are not taken into account and the value of n is then taken as the number of differences $(x_i - \eta_0)$, that are not equal to nil. If differences are found with equal absolute values, we allocate mean ranks and follow the procedure given below:

Suppose, for example, we found in Example 16.3 that $x_{10} = 261\,700$, then $|x_{10} - 300\,000| = 38\,300$ which is the same as $|x_3 - 300\,000|$. An array of the absolute differences together with the ranks that would have been allocated as well as the signs of the differences are now:

absolute difference	7 400	10 850	12 100	12 500	38 300
rank	1	2	3	4	5.5
sign of difference	–	+	–	–	+
absolute difference	38 300	69 850	90 050	119 400	130 400
rank	5.5	7	8	9	10
sign of difference	–	+	+	+	+

r_3 and r_{10} should in this case assume the values 5 and 6 (or 6 and 5). We therefore allocate the value of 5.5, which is the arithmetic mean of 5 and 6, to both. The value of the test statistic would have been the following in this case:

$$s = 2 + 5.5 + 7 + 8 + 9 + 10$$
$$= 41.5$$

and, since $s < 45$, H_0 in this case cannot be rejected.

Sequences of equal values in an array are known as **ties**. In the above situation there is one tie of **length** 2. If the ties are not too numerous and/or too long, the test procedure for the signed rank test described above can be used without any adjustments.

(c) The normal approximation of S can be used to test H_0: $\eta = \eta_0$ if n is larger than 20 (see Table 16.3).

(d) In a certain sense, the two tests for the population median η can be regarded as 'competitors' of the test for the population mean μ, which is discussed in Chapter 12, Section 5.1. However, when selecting a test one should remember that the assumptions differ with regard to the underlying population distributions and that the assumptions will in any case determine which test is to be used in the situation concerned.

3 The two-sample problem

3.1 Difference in medians (independent samples)

In the discussion of the tests for the differences in means of independent samples (Chapter 12, Section 6), we assumed that the observed samples had been taken from normal populations with equal variances or that the samples were so big that the test statistics were approximately normally distributed on the basis of the central limit theorem. The question now is how the hypothesis of equal location of two populations can be tested if the samples are small and are not necessarily taken from normal populations. The problem of comparing the players of two clubs on the basis of their positions on a ranking list can be regarded as a typical example of this situation. As in the one-sample situation we use the population median η as the measure of location to make provision for underlying distributions that could be skewed. Consider the two observed samples of sizes n_1 and n_2 (with $n_1 \leq n_2$) from populations of which the distributions are of similar shape but the medians η_1 and η_2 perhaps differ. There is a large variety of distribution-free tests, based on the joint arrangement of all $(n_1 + n_2)$ observed values, that can be used to test the hypothesis

$$H_0: \ \eta_1 = \eta_2$$

against an appropriate alternative. The oldest and best known of these tests is known as the Wilcoxon **rank sum test** and the test procedure is explained in Table 16.5. The notation $\sum_{(1)} R_i$ used in the rank sum test statistic represents the sum of the ranks of the smallest sample (sample 1) where the ranks were obtained in a joint arrangement of all $(n_1 + n_2)$ observations. This test is valid only if we assume that the samples were drawn independently from two populations with distributions of similar shape.

Table 16.5 Testing H_0: $\eta_1 = \eta_2$ on the basis of two independent random samples of sizes n_1 and n_2 ($n_1 \leq n_2$) from populations with medians η_1 en η_2 respectively

Test statistic	Distribution of test statistic under H_0	Table with critical values
$W = \sum_{(1)} R_i - \frac{n_1(n_1+1)}{2}$ with $R_1, R_2, \ldots, R_{n_1+n_2}$ the ranks of the joint sample	n small: W has an exact distribution	Table B.12
	n large: Z approximately has a n(0, 1) distribution	
$Z = \frac{W - n_1 n_2/2}{\sqrt{n_1 n_2(n_1+n_2+1)/12}}$		Table B.2
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
H_1: $\eta_1 \neq \eta_2$, that is, $\eta_1 - \eta_2 \neq 0$	$w \leq w_{n_1, n_2; \alpha/2}$ or $w \geq n_1 n_2 - w_{n_1, n_2; \alpha/2}$	$2 \min\{P(W \leq w), P(W \geq w)\}$
H_1: $\eta_1 > \eta_2$, that is, $\eta_1 - \eta_2 > 0$	$w \geq n_1 n_2 - w_{n_1, n_2; \alpha}$	$P(W \geq w)$
H_1: $\eta_1 < \eta_2$, that is, $\eta_1 - \eta_2 < 0$	$w \leq w_{n_1, n_2; \alpha}$	$P(W \leq w)$

3.1.1 Tables for the rank sum test

Under the null hypothesis the rank sum test statistic W has a discrete distribution that is symmetrically distributed around $\mu_W = (n_1 n_2)/2$. As in the cases of the sign test and the signed rank test, there are tables from which critical values for the test can be found. In Table B.12 the largest integer $w_{n_1, n_2; \alpha}$ is tabulated, for which, under H_0

$$P\left(W \leq w_{n_1, n_2; \alpha}\right) \leq \alpha$$

for different values of n_1, n_2 and α respectively, with $n_1 \leq n_2$. The value $w_{n_1, n_2; \alpha}$ is the **left critical value** of the distribution of W under H_0, while, from the symmetry of the distribution under H_0, it follows that $n_1 n_2 - w_{n_1, n_2; \alpha}$ is the **right critical value**.

Thus it follows, for example, that for $n_1 = 5$, $n_2 = 10$ and $\alpha = 0.05$, the left and right critical values are given by

$$w_{5, 10; 0.05} = 11 \text{ and } (5 \times 10) - w_{5, 10; 0.05} = 39$$

Since W was defined in terms of the sum of the ranks of the smallest sample, Table B.12 was drawn up only for $n_1 \leq n_2$. The normal approximation of the distribution of W under H_0 is already satisfactory if n_1 and n_2 are both greater than 10 (see Table 16.5).

EXAMPLE 16.4

After an inter-university tennis tournament in which four universities took part, a joint ranking of the twenty best male players is compiled. The positions occupied by the players from two of the universities are given in Table 16.6. The last column in the table reflects the ranks of the players in a joint array of the ranking positions of the players from the two universities. Let η_1 and η_2 be the median positions of the players from the two universities, and assume that the two population distributions are of similar shape. The question is: Is there a difference in the quality of the male tennis players from the two universities?

Table 16.6 Calculation table for the application of the rank sum test

Sample number	Individual number	Position on ranking list	Rank, r_i
1	1	1	1
	2	3	3
	3	10	9
	4	15	10
	5	16	11
2	6	2	2
	7	4	4
	8	5	5
	9	6	6
	10	8	7
	11	9	8

Step 1

H_0: $\eta_1 = \eta_2$
H_1: $\eta_1 \neq \eta_2$

Step 2

Select $\alpha = 0.05$. The test statistic is

$$W = \sum_{(1)} R_i - \frac{5(5+1)}{2}$$

Step 3

Reject H_0 if $w \leq w_{5,6;0.025} = 3$ or if $w \geq (5 \times 6) - w_{5,6;0.025} = 27$.
Alternatively: Reject H_0 if the p-value ≤ 0.05.

Step 4

From Table 16.6 it follows that

$$w = (1 + 3 + 9 + 10 + 11) - 15$$
$$= 34 - 15$$
$$= 19$$

Step 5

Since $3 < w < 27$, H_0 is not rejected at the 5% level of significance. The quality of the male players from the two universities therefore does not appear to differ. The p-value $= 2 \min \{P(W \leq 19,\ P(W \geq 19)\} = 2(0.2684) = 0.5368 > 0.05$. This p-value is determined from a table of the cumulative distribution function of the rank sum test statistic.

Notes

(a) If two or more values in the joint array of two samples are equal, the values form a tie. As in the signed rank test, mean ranks are allocated in such a case. If the ties are not too numerous and/or too long, no adjustment is required in the rank sum test procedure, as discussed above.

(b) A question that can justifiably be asked with respect to Example 16.4, is whether or not the two groups of tennis players (the best from each university) can be regarded as random samples of players from the universities. The answer of course is no. However, we can prove that the Wilcoxon rank sum test can be applied if the positions of all the tennis players on the joint ranking list are independent of each other. In such a case it would therefore be more correct to refer to two independent sets of values and the null hypothesis for the example concerned would then be given by

H_0: The best players from the two universities do not differ.

The relevant Wilcoxon test procedure is then used as a so-called **permutation test**.

3.2 Differences in medians (dependent samples)

In Chapter 12, Section 6, in testing of the equality of two population means we distinguished between the case of independent samples (Chapter 12, Section 6.1 to 6.3) and the case of dependent samples or paired values (Section 6.4 of Chapter 12). In Section 3.1 of this chapter we defined the rank sum test for the case of independent samples. In this section we discuss the sign test and signed rank test for dependent samples. In an investigation on the effectiveness of a certain painkiller, each patient can indicate (for instance on a five-point scale) the degree of his pain before and after treatment. In tasting tests, for example, a number of connoisseurs can indicate which of two tastes they prefer, and so

forth. As in the case of the t test discussed in Chapter 12, Section 6, for this type of problem, we are here working with the differences between paired observations D_i, but in contrast to the t test, we assume that the distribution associated with the differences is unknown. We are therefore interested in testing the null hypothesis

$$H_0: \eta_D = 0$$

against an appropriate alternative, where η_D is the median of the distribution of differences.

Thus formulated, the problem of two dependent or paired samples appears to be a special case of the one-sample problem, where we had as null hypothesis

$$H_0: \eta = \eta_0.$$

From this discussion it follows that we may apply the sign test and signed rank test, as discussed in Section 2, directly to the differences $D_1, D_2, \ldots, D_n$. Here we must remember that symmetry of the distribution of the D_i's is a prerequisite for the application of the signed rank test. The following example illustrates the test procedure:

EXAMPLE 16.5

Suppose we need to determine, as in Example 12.18, whether it takes workers less time to assemble a certain part after a special training programme. We consider the data in the example again, but assume that the assumption of normality is not justified. The periods before and after the training appear in the second and third columns of Table 16.7. The other columns contain the calculations required for applying the sign test and the signed rank test.

Table 16.7 Calculation table for applying the sign test and the signed rank test to the assembly time data in Example 12.18

Worker	Before training	After training	d_i	sign	$\lvert d_i \rvert$	r_i
1	23	17	6	(+)	6	7
2	17	14	3	(+)	3	4
3	16	12	4	(+)	4	5
4	15	13	2	(+)	2	3
5	19	12	7	(+)	7	8
6	21	20	1	(+)	1	$1\frac{1}{2}$
7	13	14	−1	(−)	1	$1\frac{1}{2}$
8	20	15	5	(+)	5	6

(a) Sign test

Step 1

$H_0: \eta_D = 0$
$H_1: \eta_D > 0$

Step 2

Select $\alpha = 0.01$. The number of positive differences, B, has a bi(8, 0.5) distribution under H_0.

Step 3

Reject H_0 if $b \geq 8 - b_{8,\,0.01} = 8$ or if the p-value ≤ 0.01.

Step 4

From Table 16.7 it follows that $b = 7$.

Step 5

Since $b < 8$, H_0 is not rejected at the 1% level of significance. The p-value can be calculated as follows from Table B.6:
$P(B \geq 7) = 1 - 0.96484 = 0.03516 > 0.01$.

(b) Signed rank test

Step 1

H_0: $\eta_D = 0$
H_1: $\eta_D > 0$

Step 2

Select $\alpha = 0.01$. The test statistic that is used, is

$$S = \sum_{(+)} R_i$$

with R_i the rank of D_i.

Step 3

Reject H_0 if $s \geq 8(8 + 1)/2 - s_{8,\,0.01} = 35$ or if the p-value ≤ 0.01.

Step 4

From Table 12.7 it follows that

$$\begin{aligned} s &= 7 + 4 + 5 + 3 + 8 + 1.5 + 6 \\ &= 34.5 \end{aligned}$$

Step 5

Since $s < 35$, H_0 is once again not rejected at a 1% level of significance. The p-value is calculated from a table of the cumulative distribution function of S as:
p-value $= P(S \geq 34.5) = 0.0117 > 0.01$.

In contrast to the conclusion drawn in Example 12.18, we now conclude, with the aid of the sign test and the signed rank test, that the training programme did not effect significant decreases in assembly times. With a larger sample of workers the sign test and the signed rank test, and also the t test used in Example 12.18, might well yield the same outcome.

3.3 Difference in spread

In Chapter 5 we defined different measures for spread in a data set. To test whether two data sets differ with regard to spread, the F test for difference in variances is studied in Chapter 12, Section 6.6. The test is based on the assumption that the observed samples are taken from normal populations. However, two data sets can be tested for significant differences in spread without making the assumption of normality. There are quite a number of distribution-free tests for this purpose. We shall discuss only one, namely the Siegel-Tukey test (named after the two people who proposed the test in 1960).

In order to apply this test, we again allocate ranks to the observations of the two samples in the following way: Take two samples of sizes $n_1 = 5$ and $n_2 = 6$. An array of the joint sample, together with the ranks allocated according to the Siegel-Tukey method, is the following:

Array of joint sample	3 (1)	5 (1)	6	9 (1)	12	13	14	18 (1)	21	25	31 (1)
Ranks according to Siegel-Tukey method	1	4	5	8	9		10	7	6	3	2

The observations of the smallest sample are identified by a (1). It is clear that this scheme alternately allocates the smallest ranks to the most extreme observed values and the highest ranks to the middle group of observations. In the case of an odd number of observations the middle observation is omitted and N decreased by one. If the ranks allocated according to this method are indicated by $R'_1, R'_2, \ldots, R'_{n_1+n_2}$, the Siegel-Tukey test amounts to the application of the Wilcoxon rank sum test to these ranks. The Siegel-Tukey test procedure is given in Table 16.8. Note that we assume that the medians of the two populations from which the two samples are taken are equal. As in the case of the rank sum test, we assume here that $n_1 \leq n_2$.

Although the test statistic of the rank sum test W and the Siegel-Tukey test statistic W' are based on ranks allocated in different ways, the two statistics have the same form, which means that the distributions of W and W' under the relevant null hypotheses must be the same. This means that we can use the tables of the rank sum test when carrying out the Siegel-Tukey test. The normal approximation of W' (see Table 16.8) is also satisfactory, as it is for the rank sum test, if n_1 and n_2 are both larger than 10.

Table 16.8 Testing for equality in the spread of two populations on the basis of two independent random samples of sizes n_1 and $n_2 (n_1 \leq n_2)$ from populations with equal medians

Test statistic	Distribution of test statistic under H_0	Table with critical values
$W' = \sum_{(1)} R'_i - \frac{n_1(n_1+1)}{2}$ with $R'_1, R'_2, \ldots, R'_{n_1+n_2}$ the rank of the joint sample	n small: W has an exact distribution	Table B.12
$Z = \frac{W' - n_1 n_2/2}{\sqrt{n_1 n_2 (n_1+n_2+1)/12}}$	n large: Z approximately has a n(0, 1) distribution	Table B.2
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
H_1: Population spreads differ	$w' \leq w_{n_1, n_2; \alpha/2}$ or $w' \geq n_1 n_2 - w_{n_1, n_2; \alpha/2}$	$2 \min \{P(W \leq w'), P(W \geq w')\}$
H_1: Spread of first population the largest	$w' \geq n_1 n_2 - w_{n_1, n_2; \alpha}$	$P(W \geq w')$
H_1: Spread of first population the smallest	$w' \leq w_{n_1, n_2; \alpha}$	$P(W \leq w')$

EXAMPLE 16.6

Suppose we need to test at a 5% level of significance whether there is a difference in the spread in the floor areas of the main buildings of residences on the East and the West Rand. A random sample of size 10 is drawn from the population of East Rand houses and, together with a random sample of sixteen West Rand houses, these produce the following data:

East Rand	100	155	300	121	126	134	183	143	105	116
West Rand	100	140	188	210	145	180	142	143		
	205	100	182	198	194	150	155	150		

We now want to use the Siegel-Tukey test to determine whether the two populations differ with respect to spread in the area of main buildings, assuming, however, that the population medians are equal.

Step 1

H_0: The spread in the areas of main buildings is the same for the East Rand and the West Rand
H_1: The spread in the areas of main buildings differs.

Step 2

Select $\alpha = 0.05$. Under H_0

$$Z = \frac{W' - (10 \times 16)/2}{\sqrt{10 \times 16(10 + 16 + 1)/12}}$$

with $W' = \sum_{(1)} R'_i - \frac{10(11)}{2}$

approximately has a n(0,1) distribution.

Step 3

Reject H_0 if $z \leq -1.96$ or if $z \geq 1.96$, or if the p-value ≤ 0.05.

Step 4

The array of the 26 values given above, together with their ranks according to the Siegel-Tukey method, is as follows (the smallest sample is again indicated by (1)):

Array	100 (1)	100	100	105 (1)	116 (1)	121 (1)	126 (1)	134 (1)	140
Rank	2	2	2	6	7	10	11	14	15
Array	142	143	143 (1)	145	150	150	155 (1)	155	180
Rank	18	$19\frac{1}{2}$	$19\frac{1}{2}$	24	$25\frac{1}{2}$	$25\frac{1}{2}$	$22\frac{1}{2}$	$22\frac{1}{2}$	21
Array	182	183 (1)	188	194	198	205	210	300 (1)	
Rank	17	16	13	12	9	8	5	4	

From the above table it now follows that

$$w' = (2 + 6 + 7 + 10 + 11 + 14 + 19\tfrac{1}{2} + 22\tfrac{1}{2} + 16 + 4) - 55$$
$$= 57$$

From this it follows that

$$z = \frac{57 - 80}{\sqrt{360}}$$
$$= -1.212$$

Step 5

Since $-1.96 < z < 1.96$, H_0 is not rejected at a 5% level of significance. There is thus no significant difference in the spread between the areas of the main buildings in the two samples. From Table B.2 it follows that the p-value $= 2P(Z \leq -1.212) = 2(0.11314) = 0.226 > 0.05$.

Note

Application of the Siegel-Tukey test necessitates that the two samples do not differ with respect to their medians. This assumption is not necessary for the application of the F test for equal variances.

4 The multi-sample problem

Suppose we need to compare with one another, the members of three different sports clubs, each of which plays a specific type of sport. The comparison is with regard to the players' positions on a mutual ranking list (see Section 1). In Example 16.4 we saw how the problem is solved by a distribution-free method if we need to compare the players from two university clubs with one another. We can now compare the three clubs in pairs by means of the Wilcoxon rank sum test, that is, Club A with Club B, Club A with C and Club B with C. The problem, however, is that such paired comparisons can make the probability of a type I error unnecessarily large. It would therefore be valuable if a distribution-free test procedure could be found that could compare all three clubs (or more clubs, if required) with one another at a preselected level of significance. The best-known distribution-free test for this problem was proposed by W.H. Kruskal and W.A. Wallis in 1952 and is commonly known as the Kruskal-Wallis test.

If $\mu_1, \mu_2, \ldots, \mu_c$ are the means of c different populations and we need to test the null hypothesis

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_c$$

the F test for the completely randomized experiment can be used, provided the observations on which the test statistic is based can be regarded as c independent samples taken from different normal populations with the same variance (see Chapter 14, Section 2). The Kruskal-Wallis test can be used as a distribution-free test for the problem if the normality assumption is not justified.

Also consider the following problem: A panel consisting of ten wine tasters is requested to judge four types of semi-sweet wine by arranging them in order of preference. The aim is to determine, on the basis of the preferences of the wine tasters, whether there is a difference in the popularity of the different

types of wine. The **Friedman test**, proposed in 1937 by M. Friedman, is the oldest known distribution-free test for this problem.

In Chapter 14, Section 3 the random block design is proposed as an analysis of variance technique for testing the null hypothesis of equal treatment means for associated samples. In this design the experimental units are divided into homogeneous blocks in order to eliminate the effect of additional sources of variation. Since the Friedman test is not dependent on the normality assumption, as is the case with the F test, it is also used as a distribution-free technique for the problem.

From the above two examples we note that in the multi-sample problem we can distinguish between

(a) the case of c independent samples, and
(b) the case of c associated samples.

In this section we shall discuss only the above two tests, namely the Kruskal-Wallis test for case (a) and the Friedman test for case (b).

4.1 The Kruskal-Wallis test (independent samples)

The Kruskal-Wallis test can be regarded as an extension of the rank sum test for two independent samples (see Section 3.1). To calculate the test statistic, a scheme similar to the one in Table 14.1 is followed, but with the following difference: Each observation in the table is replaced by its rank in the joint array of all the observations. We also consider the more common case of unequal sample sizes $h_1, h_2, \ldots, h_c$, where $\sum_{i=1}^{c} h_i = n$. Table 16.9 gives the table of ranks and the totals $T_1, T_2, \ldots, T_c$, of the ranks of the c samples.

Table 16.9 Ranks and rank totals of the c samples required for the Kruskal-Wallis test

Treatment	Rank in joint array	Totals
1	ranks of first sample	T_1
2	ranks of second sample	T_2
.	.	.
.	.	.
.	.	.
c	ranks of cth sample	T_c
		$\frac{n(n+1)}{2}$

The Kruskal-Wallis test procedure is explained in Table 16.10. Note that the Kruskal-Wallis test can be used only if the different samples come from populations with the same general shape but with possibly different medians. The χ^2 approximation of the distribution of the test statistic K under H_0 is already satisfactory if each sample size $h_1, h_2, \ldots, h_c$ is greater than 5.

Table 16.10 Testing H_0: $\eta_1 = \eta_2 = \ldots = \eta_c$ on the basis of c independent random samples from populations with the same general shape and with medians $\eta_1, \eta_2, \ldots, \eta_c$ respectively

Test statistic	Distribution of test statistic under H_0	Table with critical values
$K = \frac{12}{n(n+1)} \sum_{i=1}^{c} \frac{T_i^2}{h_i} - 3(n+1)$ with T_i = the sum of the ith sample's ranks	n relatively large: approximate $\chi^2(c-1)$	Table B.4
Alternative hypothesis	**Decision rule: Reject H_0 if**	**p-value**
H_1: At least one η_i differs from the others	$k \geq \chi^2_{c-1,\alpha}$	$P(K \geq k)$

EXAMPLE 16.7

Consider the percentages scored by 23 students who underwent training with the aid of four different textbooks, as given in Table 14.6. If we cannot safely assume that the achievements are taken from normal populations, but that they are taken from populations with the same general shape, then the null hypothesis of equal achievements for the four groups, namely

$$H_0: \eta_1 = \eta_2 = \eta_3 = \eta_4$$

can be tested with the aid of the Kruskal-Wallis procedure. Table 16.11 contains the ranks and sums of ranks of the 23 percentages, and the original observations are indicated in brackets below the ranks.

Table 16.11 Ranks of percentages obtained by 23 students trained with the aid of four different textbooks

Textbook								Total
1	$14\frac{1}{2}$	$8\frac{1}{2}$	20	22	10	$16\frac{1}{2}$		$91\frac{1}{2}$
	(72)	(65)	(81)	(83)	(67)	(73)		
2	1	7	13	5	3	2	4	35
	(48)	(63)	(71)	(55)	(52)	(49)	(54)	
3	$14\frac{1}{2}$	19	21	$16\frac{1}{2}$	12			83
	(72)	(76)	(82)	(73)	(69)			
4	$8\frac{1}{2}$	11	18	6	23			$66\frac{1}{2}$
	(65)	(68)	(75)	(61)	(92)			
								276

Step 1

H_0: $\eta_1 = \eta_2 = \eta_3 = \eta_4$
H_1: At least two medians differ

Step 2

Select $\alpha = 0.01$. The test statistic

$$K = \frac{12}{23 \times 24} \sum_{i=1}^{4} \frac{T_i^2}{h_i} - (3 \times 24)$$

approximately has a $\chi^2(3)$ distribution under H_0.

Step 3

Reject H_0 if $k \geq \chi^2_{3,\,0.01} = 11.345$ or if the p-value ≤ 0.01.

Step 4

From Table 16.11 it follows that

$$k = \frac{12}{23 \times 24} \left\{ \frac{(91.5)^2}{6} + \frac{(35)^2}{7} + \frac{(83)^2}{5} + \frac{(66.5)^2}{5} \right\} - (3 \times 24)$$

$$= 11.32$$

Step 5

Since $k < 11.345$, H_0 is only just not rejected at a 1% level of significance. Thus there does not appear to be any significant difference between the four textbooks with regard to the students' achievements. Note, however, that $k \geq \chi^2_{3,\,0.05}$, so that H_0 would be rejected at a 5% level of significance if $\alpha = 0.05$ had been selected beforehand. With the aid of a computer program the p-value is given as $P(K \geq 11.32) = 0.01012 > 0.01$.

Notes

(a) Coping with ties by allocating mean ranks is standard procedure here as well.

(b) There are tables from which critical values for the **exact distribution** of the Kruskal-Wallis test for $c = 3$ and $h_i = 1, 2, 3$ can be determined. These tables are readily available in statistical literature.

(c) The problem of **paired comparisons** of population medians by means of distribution-free techniques is beyond the scope of this book.

4.2 The Friedman test (associated samples)

In Section 3.2 we looked at the problem of paired samples and saw that the sign test and the signed rank test can be used in this case to test the hypothesis H_0: $\eta_D = 0$. The Friedman test for c samples can be regarded as an extension of the sign test for paired observations. To understand the calculation of the test statistic, we refer to Table 16.12. In contrast to the Kruskal-Wallis procedure, this table contains a separate arrangement of the observations in each block so that within each block the ranks 1, 2, ..., c are allocated (if the observations are not ranks already).

Table 16.12 Ranks and rank totals of c samples that are required for the Friedman test

Treatments	Blocks 1	2	...	h	Totals
1	ranks	ranks	...	ranks	T_1
2	of	of		of	T_2
.	first	second		hth block	.
.	block	block			.
.	.	.	.	.	.
.	.	.	.	.	.
c	.	.	.	.	T_c
					$\frac{hc(c+1)}{2}$

In practice the Friedman test is used for two types of problems, namely as a distribution-free procedure for **random block experiments**, and as a solution to the so-called **problem of h rankings**. The exposition in Table 16.12 applies in both cases and the general test procedure is also the same.

4.2.1 The Friedman test for random block experiments

The Friedman test statistic is based on the rank totals T_1, T_2, ..., T_c of the c treatments. This test can be applied in a random block experiment only if the c samples can be regarded as having been taken from populations with the same general shape. Table 16.13 contains an exposition of Friedman's general procedure for this problem. The χ^2 approximation of the test statistic Q under H_0 is already satisfactory if $hc \geq 30$.

Table 16.13 Testing H_0: $\eta_1 = \eta_2 = \ldots = \eta_c$ on the basis of c associated samples of size h each, from populations with the same general shape and with medians $\eta_1, \eta_2, \ldots, \eta_c$

Test statistic	Distribution of test statistic under H_0	Table with critical values
$Q = \frac{12}{n(c+1)} \sum_{i=1}^{c} T_i^2 - 3h(c+1)$ with $n = hc$ and T_i = sum of the ranks of the ith sample	n relatively large: Q approximately has a $\chi^2(c-1)$ distribution	Table B.4
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
H_1: At least one η_i differs from the others	$q \geq \chi^2_{c-1,\alpha}$	$P(Q \geq q)$

EXAMPLE 16.8

Suppose we want to use random sampling to compare the mean valuation of residences in the Johannesburg area, the East Rand and the West Rand with one another. To ensure that differences in valuations do not exist solely as a result of differences in the sizes of residences, a number of blocks are selected, each consisting of three residences of approximately the same floor area, one from each of the areas mentioned. The data is given in Table 16.14, as follows:

In the first section, namely (a), the building valuation of each residence is given, together with the floor area of the residence in brackets. From this we can see that the block units are homogeneous. Section (b) of Table 16.14 gives the ranks of the building valuations, as obtained by arranging the valuations blockwise. Denote the median values of the three areas by means of η_1, η_2 and η_3 respectively.

Step 1

H_0: $\eta_1 = \eta_2 = \eta_3$
H_1: The medians differ from one another.

Step 2

Select $\alpha = 0.05$
The test statistic

$$Q = \frac{12}{11 \times 3 \times 4} \sum_{i=1}^{3} T_i^2 - (3 \times 11 \times 4)$$

approximately has a $\chi^2(2)$ distribution under H_0.

Step 3

Reject H_0 if $q > \chi^2_{2,0.05} = 5.992$ or if the p-value ≤ 0.05.

Table 16.14 Building valuations (in thousand rands) for 33 residences on the Witwatersrand, together with a Friedman rank allocation

(a)

Region	Blocks 1	2	3	4	5	6	7	8	9	10	11	Total b_i	Mean $\bar{b}_i$
1. Johannesburg	320	160	195	165	300	315	120	170	120	300	210	2 375	215.9
	(387)	(160)	(163)	(134)	(183)	(190)	(166)	(212)	(175)	(183)	(214)		
2. East Rand	430	405	170	160	206	190	165	215	159	150	320	2 270	206.4
	(390)	(160)	(162)	(134)	(183)	(189)	(167)	(212)	(174)	(182)	(213)		
3. West Rand	515	185	95	185	220	160	205	240	305	240	235	2 585	235.0
	(376)	(160)	(164)	(134)	(185)	(188)	(168)	(210)	(174)	(180)	(213)		
Total, k_i	1 265	450	460	510	726	665	490	625	584	690	765	t = 7 230	
Mean, $\bar{k}_i$	421.7	150.0	153.3	170.0	242.0	221.7	163.3	208.3	194.7	230.0	255.0		$\bar{x}$ = 219.1

(b)

Region	Blocks 1	2	3	4	5	6	7	8	9	10	11	Total
Johannesburg	1	2	3	2	3	3	1	1	1	3	1	21
East Rand	2	1	2	1	1	2	2	2	2	1	3	19
West Rand	3	3	1	3	2	1	3	3	3	2	2	26
												66

Step 4

From Table 16.14 it follows that

$$q = \frac{12}{132}\left\{(21)^2 + (19)^2 + (26)^2\right\} - 132$$
$$= 2.36$$

Step 5

Since $q < 5.992$, H_0 is not rejected at a 5% level of significance and we can say that the valuations of residences in the three regions on the Witwatersrand are the same if houses of the same size are compared. With the aid of a computer program the p-value can be determined as: $P(Q \geq 2.36) = 0.30738 > 0.05$.

4.2.2 The Friedman test for the problem of h rankings

In the problem of h rankings we test for correspondence of h independent rankings of c entities in order of preference. The second example referred to at the beginning of Section 4, namely where ten wine tasters each had to judge different types of semi-sweet wines and arrange them in order of preference, is a typical example of this problem. The general test procedure of Table 16.13 applies here as well, although the premise here differs from the one given in Table 16.13: We specifically consider testing

H_0: There is no correspondence between the h rankings, that is, the entities are equally popular; as opposed to
H_1: The h rankings correspond.

If H_1 is accepted, it means that the c competing entities have significantly different rank totals and that the entities are therefore not equally 'good' or 'popular'.

EXAMPLE 16.9

Consider the example referred to above, where each of ten wine tasters has to judge four types of semi-sweet wine by arranging the types in order of preference. Table 16.15 contains the orders of preference of the ten wine tasters and the sums of the ranks of each type of wine. Since the orders of preference are given in terms of ranks, the Friedman test can be applied directly to the rank totals.

Step 1

H_0: The four types of wine are equally popular.
H_1: The four types of wine are not equally popular.

Step 2

Select $\alpha = 0.05$. Since $n > 30$,

Table 16.15 Orders of preference of ten wine tasters for four types of wine

Type of wine	Wine taster										Totals
	1	2	3	4	5	6	7	8	9	10	
A	2	2	1	4	4	2	3	4	1	4	27
B	3	3	4	3	1	3	2	2	2	1	24
C	1	1	2	1	2	1	1	1	3	2	15
D	4	4	3	2	3	4	4	3	4	3	34
											100

$$Q = \frac{12}{10 \times 4 \times 5} \sum_{i=1}^{4} T_i^2 - (3 \times 10 \times 5)$$

approximately has a $\chi^2(3)$ distribution under H_0.

Step 3

Reject H_0 if $q \geq \chi^2_{3,0.05} = 7.815$ or if the p-value ≤ 0.05.

Step 4

According to Table 16.15

$$q = \frac{12}{200}\left\{(27)^2 + (24)^2 + (15)^2 + (34)^2\right\} - 150$$
$$= 11.16$$

Step 5

Since $q \geq 7.815$, H_0 is rejected at a 5% level of significance and we can thus say that the four types of wine are not equally popular among the wine tasters. The p-value is $P(Q \geq 11.16) = 0.01089 < 0.05$.

Notes

(a) As for the Kruskal-Wallis test, there are tables of critical values for the **exact distribution** of the Friedman test statistic, based on small values of c and h.

(b) As already stated, **pairwise comparisons** are beyond the scope of this book.

(c) As usual, mean ranks are used to deal with ties. In the case of **dichotomous** data (each observation can be only 0 or 1), there are only one or two ties in the array of the observations in each block. In this case, Cochran's test can

be used to determine whether or not c treatment medians differ from one another (see also Chapter 15, Section 3.2 for a formulation in terms of a contingency table).
Let

T_i = number of ones in ith treatment (row)
K_j = number of ones in jth block (column)

The Cochran test statistic is given by

$$C = \frac{(c-1)\left\{c\sum_{i=1}^{c} T_i^2 - \left(\sum_{i=1}^{c} T_i\right)^2\right\}}{c\sum_{j=1}^{h} K_j - \sum_{j=1}^{h} K_j^2}$$

Under the null hypothesis of equal treatment medians, C approximately has a $\chi^2(c-1)$ distribution, provided n is relatively large. (See Exercise 23 at the end of the chapter for an application.)

(d) A measure of correspondence between the h rankings of the Friedman method is given by

$$W = Q/\{h(c-1)\}$$

and is known as **Kendall's coefficient of concordance**.

5 The problem of randomness

One of the most basic assumptions made in statistical inference is that the observed sample values have been obtained randomly. The researcher is often faced with the question: Does the observed sample actually meet the requirements of **randomness**? The **runs test** for randomness is very simple to apply, provided the sequence in which the observations were made, is known. In a survey on street noise, for example, a fieldworker is given the assignment of conducting interviews with the residents of a large block of flats. How can the project leader be sure that the fieldworker did in fact draw a random sample? To apply the test, the sample observations should be **dichotomized** in some way or other, that is, the observations must be divided into two categories according to a suitable criterion. In the above example the fieldworker can indicate the numbers of the flats in the sequence in which they realized (according to the random mechanism he used). Each flat number can be replaced by a 0 or a 1, depending on whether the number is smaller or greater then the median number. In some cases the observations are naturally dichotomous. Examples here are: tossing a coin (heads or tails), or observing people's attitudes on a contentious issue (support or oppose). The stochastic variable used in the runs test is the **number of runs** L, where a run is defined as follows:

In a dichotomous data set with n_1 symbols of the one type and n_2 of the other type and in which the values are recorded in the sequence they were observed, a consecutive group of the same symbols indicates a **run**.

Take the following dichotomous data set, for example:

$$++-+-++++--+-+-$$

Here $n_1 = 9$, $n_2 = 6$ and the observed number of runs $\ell = 10$. It is clear that for a **random** dichotomous data set recorded in the sequence in which it was collected, there should be neither too many nor too few runs. The runs test is thus a test for

H_0: Observed sample is random

against

H_1: Observed sample is not random.

The full test procedure is explained in the usual way in Table 16.16.

Table 16.16 Testing the null hypothesis of randomness by means of the runs test

Test statistic	**Distribution of test statistic under H_0**	**Table with critical values**
$Z = \frac{L - \mu_L}{\sigma_L}$ with $\mu_L = \frac{2n_1n_2}{n_1+n_2} + 1$ $\sigma_L^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$	n_1 and n_2 large: Z approximately has a n(0, 1) distribution	Table B.2
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
H_1: Observed sample is not random.	$z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$	$2P(Z \geq \lvert z \rvert)$

EXAMPLE 16.10

Suppose the fieldworker referred to above has drawn a 'random' sample of 30 flat units from a large block of flats. The numbers of the units were recorded in sequence as follows:

41	61	42	82	47	68	44	59	21	52
23	87	26	54	84	48	22	29	56	11
51	13	58	81	67	28	89	62	19	16

The array of these values is as follows:

11	13	16	19	21	22	23	26	28	29
41	42	44	47	48	51	52	54	56	58
29	61	62	67	68	81	82	84	87	89

The median is given by

$$me = \frac{48 + 51}{2}$$
$$= 49.5$$

Indicate the values in the raw data set that are smaller than 49.5 with zeros, and the values greater than 49.5 with ones:

0	1	0	1	0	1	0	1	0	1
0	1	0	1	1	0	0	0	1	0
1	0	1	1	1	0	1	1	0	0

From this

$$n_1 = 15, \quad n_2 = 15, \quad \ell = 23$$
$$\mu_L = 16.0, \quad \sigma_L = 2.691$$

The rest of the test procedure is now carried out in the usual steps:

Step 1

H_0: Sample was drawn randomly.
H_1: Sample was not drawn randomly.

Step 2

Select $\alpha = 0.05$. Under H_0

$$Z = \frac{L - 16.0}{2.691}$$

approximately has a n(0, 1) distribution.

Step 3

Reject H_0 if $z \leq -1.96$ or if $z \geq 1.96$, or if the p-value ≤ 0.05.

Step 4

$$z = \frac{23 - 16}{2.691}$$
$$= 2.60$$

Step 5

Since $z \geq 1.96$, it follows that H_0 is rejected at a 5% level of significance. It therefore seems as though the fieldworker did not draw the sample randomly (note

that none of the flats on the third and seventh floors were included!). The p-value = $2P(Z \geq 2.60) = 2(0.00466) < 0.05$.

Note

There are tables of the exact distribution of L under H_0 for small values of n_1 and n_2. With the aid of these tables H_0 can be tested against H_1 for small samples. The normal approximation is already satisfactory for n_1 and $n_2 > 10$.

6 Rank correlation

In Chapter 6 we saw that, when observing bivariate data sets (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), the relation between the two characteristics can be described by means of some sort of curve, for example a straight line or a power curve. We also saw that the strength of the linear relation between x and y is measured by means of the correlation coefficient r. In Chapter 13 we discussed statistical inference with regard to a linear relation between two population characteristics and tested, inter alia, hypotheses with regard to the population correlation coefficient. Since the correlation coefficient r is a measure of linear dependence that is usually based on quantitative data, there is a need for a measure of general dependence (monotone dependence) between two characteristics that can be measured only on an ordinal scale. For instance, it might be important to a students' tennis club to determine the correlation between the quality of the game and the academic achievements of tennis players. The quality of the game is determined by the club's ranking list (see Example 16.11 below).

To meet this requirement, measures of rank correlation were developed. Of these, one of the best known is Spearman's rank correlation coefficient.

Spearman's rank correlation coefficient is used as a measure of the strength of the general relation between two characteristics X and Y and is calculated from

$$R_S = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n^3 - n},$$

where D_i = (rank of X_i) − (rank of Y_i) in **separate** arrangements of the X and Y values.

Table 16.17 shows how R_S can be used in the case of a bivariate sample of size $n > 10$ to test whether or not the rank correlation between two population characteristics differs from 0. Under

H_0: No rank correlation between two population characteristics

R_S is distributed symmetrically around zero and, as in the case of the correlation coefficient r, it can assume only values between −1 and 1.

Table 16.17 Testing H_0: No monotone relation between X and Y, based on n pairs of ranks

Test statistic	Distribution of test statistic under H_0	Table with critical values
$Z = R_S\sqrt{n-1}$ where $R_S = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{(n^3 - n)}$ and D_i = difference in ranks of the ith entity	n greater than 10: Z approximately has a n(0, 1) distribution	Table B.2
Alternative hypothesis	**Decision rule: Reject H_0 if**	***p*-value**
H_1: Monotone relationship between X and Y	$z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$	$2P(Z \geq \lvert z \rvert)$
H_1: Monotone increasing relationship between X and Y	$z \geq z_{\alpha}$	$P(Z \geq z)$
H_1: Monotone decreasing relationship between X and Y	$z \leq -z_{\alpha}$	$P(Z \leq z)$

We now explain the calculation of R_S and the testing of the significance of a value thus obtained on the basis of the following example.

EXAMPLE 16.11

We have already referred to the establishment of the correlation between the quality of the game of tennis players at a certain university and their academic achievements. Table 16.18 now indicates the ranks of 11 students according to their positions on the university club's ranking list. Since the eleven students are from different class groups at the university, their academic achievements cannot be compared directly. The technique of standardization (see Chapter 5, Section 4) was used to determine each student's standardized achievement score and the figures are indicated in column 3 of Table 16.18. This table also shows the required operations for calculating Spearman's rank correlation coefficient on the basis of the ranks of quality of the game, on the one hand, and academic achievement on the other. We shall now test whether tennis performance and academic achievement are related.

Step 1

H_0: There is no monotone relationship between tennis performance and academic achievement.

H_1: There is a monotone relationship between tennis performance and academic achievement.

Table 16.18 Calculation table for Spearman's rank correlation coefficient

Student number	Rank of quality of game	Standardised achievement	Rank of achievement	d_i	d_i^2
1	1	2.41	2	–1	1
2	3	2.05	3	0	0
3	9	–1.87	9	0	0
4	10	–1.53	8	2	4
5	11	–0.56	7	4	16
6	2	0.48	5	–3	9
7	4	1.46	4	0	0
8	5	2.89	1	4	16
9	6	–2.01	10	–4	16
10	7	0.47	6	1	1
11	8	–3.19	11	–3	9
					72

Step 2

Select $\alpha = 0.05$. Under H_0

$$Z = R_S\sqrt{10} \text{ with } R_S = 1 - \frac{6\sum_{i=1}^{11} D_i^2}{(11^3 - 11)}$$

approximately has a n(0, 1) distribution.

Step 3

Reject H_0 if $z \leq -1.96$ or if $z \geq 1.96$ or if the p-value ≤ 0.05.

Step 4

$$r_S = 1 - \frac{6\sum_{i=1}^{11} d_i^2}{(n^3 - n)}$$

$$= 1 - \frac{(6 \times 72)}{(11^3 - 11)} \quad \text{(From Table 16.18)}$$

$$= 0.6727$$

so that

$$z = 0.6727\sqrt{10}$$

$$= 2.1273$$

Step 5

Since $z \geq 1.96$, H_0 is rejected at a 5% level of significance. The tennis performance and academic achievement of tennis players at the university concerned are thus shown to be interdependent. The positive correlation obtained, points to a tendency of good players being academic achievers. The p-value was calculated from Table B.2 as $2P(Z \geq 2.1273) = 2(0.01659) = 0.033 < 0.05$.

Note

Ranks should always be allocated in such a way that there can be no confusion when the rank correlation coefficient is interpreted. In the above example the lowest ranks were used for the best tennis performance and for the best academic achievement, although, in the usual array from small to large, the latter achievement would have had the highest rank. If ranks were allocated in this way, it would be difficult to interpret $R_S = -0.6727$ as a positive relation between quality of game and academic achievement!

EXAMPLE 16.12

In a survey on the preferences of holidaymakers at a public holiday resort, the management asked a number of middle-aged and a number of newly-wed couples to express their preferences with regard to various facilities. These orders of preference appear as ranks in Table 16.19, in which a rank of 1 indicates the highest preference, a rank of 2 the second highest preference, and so forth. We now need to calculate Spearman's rank correlation coefficient and then test whether there is correspondence (positive correlation) between the choices of the middle-aged and the newly-wed couples. The 'observed values' are ranks, so that the differences $d_1, d_2, \ldots, d_9$ can be determined directly as the differences between the observed orders of preference.

Table 16.19 Calculation of Spearman's rank correlation coefficient on the basis of the respective orders of preference of middle-aged and newly wed couples for facilities at holiday resorts

Facility	Ranks		d_i	d_i^2
	Middle-aged couples	Newly-weds		
Tennis courts	3	7	–4	16
A la carte restaurant	7	5	2	4
Swimming pools	1	1	0	0
Playground for young children	5	3	2	4
Mini-golf course	4	6	–2	4
Fast-food restaurant	2	2	0	0
Supertube	6	4	2	4
Bowling green	9	9	0	0
Jacuzzis	8	8	0	0
				32

Step 1

H_0: The orders of preference of the middle-aged and the newly-wed couples are uncorrelated.
H_1: The orders of preference of the middle-aged and the newly-wed couples are positively correlated.

Step 2

Select $\alpha = 0.05$. Under H_0

$$Z = R_S\sqrt{8}, \text{ with } R_S = 1 - \frac{6\sum d_i^2}{(9^3 - 9)}$$

approximately has a n(0, 1) distribution.

Step 3

Reject H_0 if $z \geq 1.645$ or if the p-value ≤ 0.05.

Step 4

From Table 16.19 it follows that

$$r_S = 1 - \frac{6 \times 32}{720}$$
$$= 0.73$$

From this it follows that

$$z = 0.73\sqrt{8}$$
$$\approx 2.06$$

Step 5

Since $z > 1.645$, H_0 is rejected at a 5% level of significance (the p-value = $P(Z \geq 2.06) = 0.0197$). There is thus a high degree of positive correlation between the preferences of middle-aged couples and newly-wed couples for facilities at a holiday resort.

Notes

(a) We should emphasize that R_S is not simply a measure of linear relation between two population characteristics, but a general measure of any **monotone relation** (for example an exponential relation) between the characteristics.

(b) Kendall's **rank correlation coefficient** is another equally well-known rank correlation coefficient and can be found in the statistical literature.

(c) There are tables of critical values of the **exact distribution** of R_S under H_0 for small values of n.

(d) It can be proved that R_S is actually nothing other than an ordinary linear correlation coefficient R (defined in Chapters 6 and 13), as calculated for the ranks of the X and Y values.

7 Estimation

There is a close relationship between the development of estimation theory and testing theory, as discussed in Chapters 11 and 12. In statistical inference on the population mean μ, for example, we see that the point estimate $\bar{X}$ is used to form the pivotal quantity $Z = \dfrac{(\bar{X} - \mu)}{S/\sqrt{n}}$. From this pivotal quantity and its distribution, a confidence interval for μ and also a test for H_0: $\mu = \mu_0$ was constructed. The logical question now is: To what extent is it possible to develop a 'distribution-free' estimation theory? Since, in distribution-free methods, we usually work with the ranks of the observations and not with the observations themselves, the relation between estimation theory and testing theory cannot be similarly formed via a pivotal quantity, as in the traditional tests. It is, however, possible to find point estimates and confidence intervals for the population parameters involved in a number of cases, on the basis of the relevant test statistics and their distributions.

7.1 Point estimation of the median η

In the case of a skewed distribution the median η is a more representative measure of location than the mean μ.

The usual estimate of the median η is given by the sample median *me*.

EXAMPLE 16.13

Consider again the annual income of 30 persons working for a large firm, as given in Table 16.2. We have already seen that $\bar{x} = 79.72$ thousand rand for the values. Since 21 of the 30 values lie below 79.72, it follows that $\bar{x}$ is not a very representative measure of location. The median value has already been determined as $me = 56.8$ thousand rand in Example 16.1. The point estimate of η is therefore 56.8 thousand rand.

7.2 Interval estimation of the median η

To find a confidence interval for η in the case of a skewed distribution, we make use of both the critical values of the sign test and the array of the observed sample $x_1, x_2, \ldots, x_n$.

If $x_1, x_2, \ldots, x_n$ is an observed sample from a population with median η, a 95% confidence interval for η is given by

$$(x_\ell, x_h)$$

Here x_ℓ and x_h are, respectively, the x values in the $(b_{n,\,0.025} + 1)$th and the $(n - b_{n,\,0.025})$th positions of the array of x values, and $b_{n,\,0.025}$ is the left critical value of the sign test.

EXAMPLE 16.14

Consider again the annual income of the thirty persons as given in Table 16.2. The left critical value of the sign test is given by $b_{30,\,0.025} = 9$ so that

$$\begin{aligned} x_\ell &= \text{value of the } (9 + 1)\text{th observation} \\ &= \text{value of the 10th observation} \\ &= 32.6 \end{aligned}$$

Similarly,

$$\begin{aligned} x_h &= \text{value of the } (30 - 9)\text{th observation} \\ &= \text{value of the 21st observation} \\ &= 78.8 \end{aligned}$$

The 95% confidence interval for η is therefore given (in thousand rands) by

(32.6, 78.8)

Notes

(a) If n is so large that the available tables of the sign test are inadequate for determining the interval (x_ℓ, x_h), the values x_ℓ and x_h can be determined as follows from the normal approximation of the sign test:

$$\left(x_\ell, x_h\right) \text{ with}$$

$$\ell = \frac{(n+1)}{2} - \sqrt{\frac{n}{4}}\, z_{0.025} \text{ and}$$

$$h = \frac{(n+1)}{2} + \sqrt{\frac{n}{4}}\, z_{0.025}$$

(b) Point estimation and interval estimation for the difference between the population medians η_1 and η_2 is generally known and can be found in the statistical literature.

(c) It can be shown that for symmetrical distributions with 'heavy tails', that is, distributions with a relatively large area below the end sections of the population curve, the arithmetic mean is less effective than the median. In such cases the median is therefore a better measure of the location of the distribution than the arithmetic mean.

8 Conclusion

8.1 Further distribution-free tests

As stated in the introduction, there are several distribution-free tests for most of the problems discussed in this chapter. The definitions of these tests are usually very similar to the tests we have discussed, but with the following difference: Almost all the tests in this chapter are based on the ranks of the observations, while other tests are based mainly on rank functions. Thus, for example, the so-called **Van der Waerden test** for the difference between two population medians is based on the test statistic

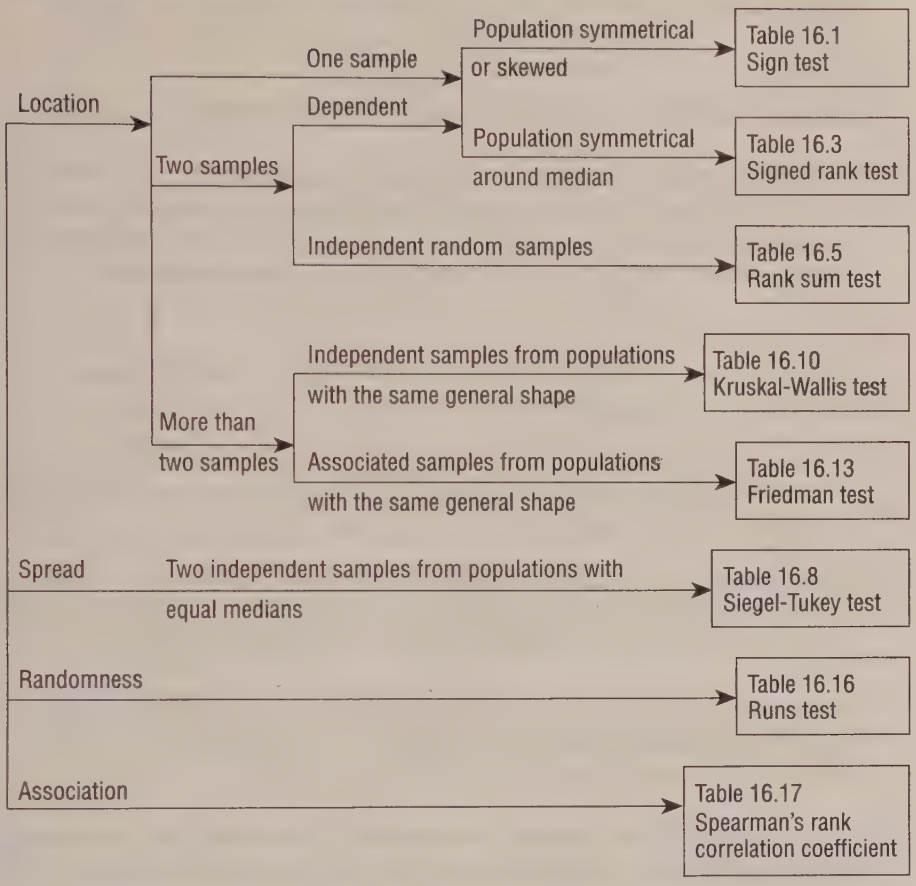
$$V = \sum_{(1)} \Phi^{-1}\left(\frac{R_i}{n_1 + n_2 + 1}\right)$$

where Φ^{-1} is the inverse distribution function of the n(0, 1) distribution and $R_1, R_2, \ldots, R_{n_1+n_2}$ has the same meaning as in Wilcoxon's rank sum test. Tests like the Van der Waerden test and the rank sum test can be used to test the equality of two population medians. Prior knowledge of the general shapes of the underlying population distributions enables the researcher to select the 'best' rank test for the specific situation.

8.2 Comparison of tests

In our discussion of the different distribution-free tests we referred to well-known tests for the problems in previous chapters. We also emphasized that the distribution-free tests were developed for situations in which the assumption of normality is not realistic, or where the data is nominal or ordinal. The comparison of different statistical tests for the same problem can be made with the aid of the concept of **relative efficiency**. The relative efficiency of one test relative to another gives an indication of the sample sizes, relatively speaking, required for each test to obtain the same power to distinguish between H_0 and H_1. If test A, for example, has a relative efficiency of 0.75 relative to test B, it means roughly that test B, on the basis of a sample of size 75, has the same power to correctly reject H_0 as does test A, based on a sample of size 100. A comparison between the distribution-free tests discussed in this chapter and the traditional tests (based on the assumption of normality) discussed in Chapters 12, 13 and 14, is beyond the scope of this book. However, it is important to remember that in cases where the assumption of normality is justifiable, the traditional tests for each problem are preferred to the distribution-free tests. If the assumption of normality cannot be justified, however, and/or if the data is of a nominal or ordinal nature, distribution-free tests are preferred.

Guide for distribution-free tests



Exercises

1 Suppose we cannot assume that the selling prices of houses for the period October to December 1991 are normally distributed and that we want to test the hypothesis

H_0: $\eta = 150\,000$ rand

against

H_1: $\eta < 150\,000$ rand

where η is the median selling price. Use the values of the following residences for the sample (Appendix A, Data Set 1):

Durban:	Numbers 101, 102, ..., 110
Johannesburg:	Numbers 101, 102, ..., 110
Cape Town:	Numbers 101, 102, ..., 110
Pretoria:	Numbers 101, 102, ..., 110.

Use the sign test to test the above hypothesis at a 5% level of significance.

2 A psychologist conducts the following experiment: Twenty people are asked to choose any integer between 0 and 9 and to note it on a piece of paper. He suspects that the median of these chosen numbers will be greater than 4. The twenty people noted the following integers (given as an array below):

1, 1, 2, 2, 3, 4, 5, 5, 5, 6, 7, 7, 7, 7, 7, 8, 8, 9, 9, 9.

Test if the psychologist's suspicions were correct by using the signed rank test at a 5% level of significance.

3 If we can assume that the distribution of the selling prices in Exercise 1 is symmetrical, what could we conclude if the normal approximation of the signed rank test were used?

4 Suppose the assumption of symmetry in Exercise 2 is debatable, and we therefore decide to solve this problem with the sign test instead. Test the psychologist's assumptions once again at a 5% level of significance.

5 Use the data from Chapter 12, Exercise 20 and the rank sum test to determine if the median braking time of the 1 500 cm^3 motor cars is significantly less than that of the 2 000 cm^3 motor cars.

6 With reference to the problem in Chapter 12, Example 15, test if the agriculturist's assumption was correct by using the rank sum test at a 5% level of significance. Use the samples used in Chapter 12, Example 15.

7 Use both the sign and signed rank tests on the data of Chapter 12, Example 17 to determine whether the number of landings of the Boeing 737 with tyres of type A are less than with those of type B ($\alpha = 0.05$).

8 Use the data in Chapter 12, Exercises 3 and 20 and test whether there is a difference in spread between the braking times of the two motor cars at a 10% level of significance by using the Siegel-Tukey test.

9 Use the Siegel-Tukey test and the data given in Chapter 12, Example 16 concerning the amount of triglyceride of coronary patients and marathon athletes, and test whether the two groups' triglyceride levels differ in spread at a 5% level of significance.

10 Use the Kruskal-Wallis test and the data in Chapter 14, Exercise 1 to test whether the four companies' 1 200 cm^3 motor car models differ with respect to petrol consumption at a 5% level of significance.

11 Use the Kruskal-Wallis test at a 5% level of significance and the data in Chapter 14, Exercise 3 to determine if the five cities differ with respect to the number of grills that are sold.

12 Use the Friedman test to answer the first question in Chapter 14, Exercise 4.

13 Twelve first-year students, all with the same four subjects, are asked to rank these subjects according to preference. The following results were obtained:

	Student											
Subject	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Business Economics	4	2	1	4	1	3	4	2	3	1	4	4
Economics	3	4	2	2	3	2	3	4	4	2	3	2
Accountancy	2	1	4	3	2	1	2	3	1	3	1	1
Statistics	1	3	3	1	4	4	1	1	2	4	2	3

Test if the four subjects are equally popular at a 5% level of significance.

14 Calculate Spearman's rank correlation coefficient between the house valuation, x, and the selling price, y, for the data given in Table 13.1 for
(a) Sample 1,
(b) Sample 4.
Test for a significant relationship ($\alpha = 0.01$) between the two characteristics in (a) and for a positive relationship in (b).

15 Eight prominent economists are asked to rank, in order of importance, the five most important bottle-necks in South Africa's economic growth. The results are tabulated below. Do the five bottle-necks differ with respect to importance? Use a 5% level of significance.

	Economist							
Bottle-neck	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Inflation	1	2	3	5	1	2	1	2
Training	2	1	1	2	3	1	2	3
Unemployment	4	5	2	1	2	5	3	1
World economy	3	3	4	3	5	3	4	5
Political climate	5	4	5	4	4	4	5	4

16 In a study to determine whether students from four provinces in South Africa adapt equally well to student life at the University of Pretoria, 20 first year students were selected randomly, five from each province, and ranked according to their ability to adapt as determined by a student counsellor. What conclusion can be reached at a 5% level of significance?

Province	Ranks				
A	2	6	7	3	13
B	10	12	1	5	14
C	4	15	18	20	17
D	8	11	16	9	19

17 Suppose the twenty students in Exercise 16 were divided into five groups which are homogeneous with respect to IQ. The five columns of the table, for the exercise mentioned, therefore represent the IQ groups. What conclusion can now be reached at a 5% level of significance?

18 Use the Siegel-Tukey test to determine whether the two data sets in Chapter 12, Exercise 18 differ with respect to spread at a 5% level of significance.

19 Use the rank sum test and the data in Chapter 12, Exercise 18 to test whether the two recovery procedures A and B differ.

20 Using the 10 petrol consumption figures in Chapter 11, Exercise 1, determine
(a) a point estimate, and
(b) a 95% confidence interval for the population median η.

21 Using the samples in Exercise 1, determine
(a) point estimates, and
(b) 95% confidence intervals for the median selling price of the houses for each of the four cities for the period October to December 1991.

22 Suppose the psychologist in Exercise 2 asks the same group of 20 people to choose an integer between 0 and 100. The following 20 integers were obtained (expressed as an array):

7, 7, 10, 20, 25, 25, 33, 36, 49, 50,
50, 50, 75, 80, 90, 97, 98, 99, 99, 100.

Determine:
(a) a point estimate, and
(b) a 99% confidence interval for the median, η, of the integers chosen.

23 The results of the twelve students in Exercise 13 are recorded at the end of their first year. 'Fail' and 'pass' are indicated with a 0 and 1 in the table. Do the subjects differ from each other with respect to the students' achievements? ($\alpha = 0.05$) (refer to Remark (c) at the end of Section 4 above).

Subject	Student 1	2	3	4	5	6	7	8	9	10	11	12
Business economics	0	1	0	1	1	1	0	1	1	0	1	0
Economics	1	0	0	1	0	1	1	1	1	0	0	1
Accountancy	1	1	0	1	1	1	0	1	1	0	1	1
Statistics	1	1	0	1	0	0	1	0	1	0	1	1

24 (a) Use the runs test at a 5% level of significance to determine whether the first fifty selling prices of houses in Johannesburg (Appendix A, Data Set 1) were indeed random.

(b) Repeat (a) for each of the other three cities given in the data set.

25 The Weather Bureau needs to determine an interval estimate for π, the ratio of days with more than half a day's sunshine, at a particular weather station. The period to be studied is May to August. Because this study has to be completed within a month, they decide to only use July's data. 'Sunshine' and 'cloudy' for more than half the day are indicated with S and C in the table. Test whether the observations are random ($\alpha = 0.05$) and comment on the applicability of the confidence interval for π given in Chapter 11, Section 3.3.

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Weather	S	S	S	S	S	S	C	C	C	C	C	C	S	C	S	S

Day	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Weather	S	C	C	C	C	S	C	S	S	S	C	C	C	S	S

Appendix A

Data Set 1

This data set consists of a sample of 2 059 houses sold during the fourth quarter of 1991 in four cities and for which loans were registered by a bank. The number of houses in each city is: Durban – 376, Johannesburg – 843, Pretoria – 434 and Cape Town – 406. The following information was observed for each house:

No = The observation number.
City = The city
- DBN = Durban
- JHB = Johannesburg
- PTA = Pretoria
- CT = Cape Town.

Price = Selling price (in R1 000).
Value = Value (in R1 000).
Area = Area of house in square metres.
Size = Size of plot
- 1 = 750 m^2 and smaller
- 2 = Bigger than 750 m^2 but smaller than or equal to 1 500 m^2
- 3 = Bigger than 1 500 m^2.

Age = Age of house
- O = Old (built before 1985)
- N = New (built during 1985 or thereafter).

PSM = Price per square metre = $\frac{\text{Selling price}}{\text{Area of house}}$.

Note that the size of the plot was not taken into account in determining PSM.

No	City	Pnce	Value	Area	Size	Age	PSM
1	DBN	130.0	111.0	186	2	O	698.92
2	DBN	78.0	75.0	122	1	N	639.34
3	DBN	142.0	122.0	150	3	O	946.67
4	DBN	130.0	117.0	166	1	O	783.13
5	DBN	120.0	115.0	223	2	O	538.12
6	DBN	165.0	270.0	316	2	O	522.15
7	DBN	150.0	132.0	215	2	O	688.07
8	DBN	145.0	132.0	225	1	O	644.44
9	DBN	56.5	53.5	89	1	O	634.83
10	DBN	155.0	147.0	300	3	O	516.67
11	DBN	175.0	165.0	185	3	O	945.95
12	DBN	85.0	83.0	97	1	O	876.29
13	DBN	87.0	82.0	251	2	O	346.61
14	DBN	90.0	200.0	207	1	O	434.78
15	DBN	68.0	66.0	113	1	O	601.77
16	DBN	100.0	95.0	193	2	O	518.13
17	DBN	40.0	38.0	88	1	O	454.55
18	DBN	107.0	85.0	154	1	O	694.81
19	DBN	158.0	125.0	152	2	O	1 039.47
20	DBN	140.0	133.0	148	1	O	945.95
21	DBN	145.0	140.0	152	2	O	953.95
22	DBN	120.0	115.0	158	3	O	759.49
23	DBN	250.0	235.0	269	3	O	929.37
24	DBN	104.0	94.0	87	3	O	1 195.40
25	DBN	227.5	220.0	273	3	O	833.33
26	DBN	90.0	145.0	180	3	O	500.00
27	DBN	144.0	119.0	184	2	O	782.61
28	DBN	139.0	120.0	206	2	O	674.76
29	DBN	82.5	90.0	139	2	O	593.53
30	DBN	176.0	172.0	207	3	O	850.24
31	DBN	110.0	105.0	158	2	N	696.20
32	DBN	215.0	280.0	246	2	O	873.98
33	DBN	168.0	146.0	193	2	O	860.10
34	DBN	138.0	113.0	195	2	O	707.69
35	DBN	126.0	98.0	128	3	O	984.37
36	DBN	133.0	108.0	139	3	O	956.83
37	DBN	170.0	165.0	284	1	N	598.59
38	DBN	128.0	123.0	178	3	O	739.88
39	DBN	102.0	98.0	84	2	O	1 214.29
40	DBN	265.0	262.0	250	3	N	1 060.00
41	DBN	110.0	102.0	107	3	O	1 028.04
42	DBN	153.0	137.0	209	3	O	732.06
43	DBN	255.0	240.0	245	3	N	1 040.82
44	DBN	78.0	72.0	84	1	O	928.57
45	DBN	145.0	136.0	182	3	O	796.70
46	DBN	125.0	167.0	234	2	O	534.19
47	DBN	185.0	170.0	248	3	O	745.97
48	DBN	85.0	55.0	135	2	O	629.63
49	DBN	140.0	130.0	167	2	O	838.32
50	DBN	92.0	80.0	102	2	N	901.96

No	City	Price	Value	Area	Size	Age	PSM
51	DBN	162.0	157.0	131	1	N	1 236.64
52	DBN	170.0	145.0	235	3	O	723.40
53	DBN	153.0	134.0	163	2	O	938.65
54	DBN	250.5	231.0	321	2	O	780.37
55	DBN	185.0	185.0	203	3	N	911.33
56	DBN	65.0	84.0	98	1	O	663.27
57	DBN	222.6	205.0	195	3	N	1 153.37
58	DBN	110.0	108.0	98	2	N	1 122.45
59	DBN	120.0	95.0	219	3	O	547.95
60	DBN	106.0	87.0	147	3	O	721.09
61	DBN	84.0	122.0	133	3	O	631.58
62	DBN	110.0	100.0	173	2	O	635.84
63	DBN	190.0	187.0	240	3	N	791.67
64	DBN	162.5	157.0	281	3	O	578.29
65	DBN	134.6	133.5	117	2	N	1 150.43
66	DBN	180.0	172.0	214	2	O	841.12
67	DBN	143.0	123.0	192	2	O	744.79
68	DBN	220.0	155.0	297	3	O	740.74
69	DBN	130.9	129.4	119	2	N	1 100.00
70	DBN	265.0	240.0	350	2	O	757.14
71	DBN	225.0	182.0	177	3	O	1 271.19
72	DBN	161.5	142.0	181	2	O	892.27
73	DBN	245.0	212.0	236	3	O	1 038.14
74	DBN	210.0	205.0	239	2	N	878.66
75	DBN	100.0	190.0	288	2	O	347.22
76	DBN	315.0	305.0	317	2	O	993.69
77	DBN	365.0	330.0	265	3	O	1 377.36
78	DBN	68.0	68.0	137	1	O	496.35
79	DBN	135.0	120.0	135	2	O	1 000.00
80	DBN	145.0	143.0	195	3	O	743.59
81	DBN	282.0	275.0	227	2	O	1 242.29
82	DBN	155.0	149.0	211	2	O	734.60
83	DBN	270.0	263.0	203	2	O	1 330.05
84	DBN	100.0	99.0	162	1	O	617.28
85	DBN	92.0	92.0	105	2	O	876.19
86	DBN	175.0	175.0	208	3	N	841.35
87	DBN	90.0	88.0	204	2	O	441.18
88	DBN	75.0	72.0	91	2	O	824.18
89	DBN	173.0	164.0	166	1	O	1 042.17
90	DBN	148.0	122.0	146	2	O	1 013.70
91	DBN	130.0	155.0	108	2	O	1 203.70
92	DBN	68.2	68.0	82	1	O	831.71
93	DBN	230.0	226.0	181	2	O	1 270.72
94	DBN	173.0	171.0	122	1	N	1 418.03
95	DBN	140.0	168.0	132	2	O	1 060.61
96	DBN	139.0	109.0	166	1	O	837.35
97	DBN	65.5	65.5	85	1	O	770.59
98	DBN	130.0	122.0	181	1	N	718.23
99	DBN	100.0	95.0	166	1	O	602.41
100	DBN	285.0	220.0	314	2	O	907.64

No	City	Price	Value	Area	Size	Age	PSM
101	DBN	144.0	109.0	162	3	O	888.89
102	DBN	235.0	235.0	246	3	N	955.28
103	DBN	123.0	105.0	95	1	O	1 294.74
104	DBN	102.4	84.0	105	3	N	975.24
105	DBN	150.0	123.0	263	1	O	570.34
106	DBN	185.0	159.0	204	1	O	906.86
107	DBN	110.0	105.0	220	3	O	500.00
108	DBN	286.1	211.0	228	2	N	1 254.82
109	DBN	235.0	215.0	208	1	O	1 129.81
110	DBN	122.0	100.0	130	1	O	938.46
111	DBN	60.0	56.0	97	1	O	618.56
112	DBN	250.0	225.0	280	3	O	892.86
113	DBN	120.0	103.0	133	2	O	902.26
114	DBN	191.5	166.5	252	3	O	759.92
115	DBN	70.0	68.0	126	3	O	555.56
116	DBN	144.0	135.0	191	2	O	753.93
117	DBN	138.0	118.0	157	2	O	878.98
118	DBN	207.0	165.0	202	3	O	1 024.75
119	DBN	265.0	240.0	350	2	O	757.14
120	DBN	100.0	130.0	178	3	O	561.80
121	DBN	114.4	114.4	100	2	N	1 144.00
122	DBN	135.0	130.0	165	2	O	818.18
123	DBN	225.0	182.0	177	3	O	1 271.19
124	DBN	214.3	208.0	206	2	O	1 040.29
125	DBN	183.0	173.0	139	2	O	1 316.55
126	DBN	98.0	97.0	142	1	O	690.14
127	DBN	215.0	173.0	138	2	O	1 557.97
128	DBN	275.0	255.0	229	2	O	1 200.87
129	DBN	135.0	130.0	163	2	O	828.22
130	DBN	141.0	121.0	184	2	O	766.30
131	DBN	167.0	160.0	203	2	O	822.66
132	DBN	84.0	80.0	90	1	O	933.33
133	DBN	90.0	85.0	135	2	O	666.67
134	DBN	120.0	107.0	139	1	O	863.31
135	DBN	83.0	70.0	99	1	O	838.38
136	DBN	177.8	166.0	99	2	N	1 795.96
137	DBN	68.0	110.0	154	2	O	441.56
138	DBN	80.0	85.0	143	2	O	419.58
139	DBN	147.0	142.0	148	3	O	993.24
140	DBN	110.0	100.0	167	2	O	658.68
141	DBN	85.0	83.0	97	1	O	876.29
142	DBN	82.6	78.0	123	2	O	671.54
143	DBN	105.0	100.0	162	3	O	648.15
144	DBN	230.0	202.0	245	1	O	938.78
145	DBN	160.0	135.0	285	3	O	561.40
146	DBN	117.5	115.5	82	2	O	1 432.93
147	DBN	143.0	128.0	132	1	O	1 083.33
148	DBN	165.0	137.0	217	2	O	760.37
149	DBN	55.0	52.0	84	1	O	654.76
150	DBN	85.5	85.5	97	1	O	881.44

No	City	Price	Value	Area	Size	Age	PSM
151	DBN	95.0	86.0	95	1	O	1 000.00
152	DBN	75.2	69.2	128	1	O	587.50
153	DBN	120.0	112.0	197	1	O	609.14
154	DBN	62.0	62.0	82	1	O	756.10
155	DBN	116.0	91.0	128	2	O	906.25
156	DBN	128.0	103.0	171	3	O	748.54
157	DBN	145.0	125.0	168	2	O	863.10
158	DBN	115.0	135.0	175	3	O	657.14
159	DBN	100.0	93.0	213	2	O	469.48
160	DBN	100.0	95.0	161	2	O	621.12
161	DBN	310.0	285.0	281	3	O	1 103.20
162	DBN	115.0	110.0	108	2	N	1 064.81
163	DBN	150.0	145.0	136	2	O	1 102.94
164	DBN	118.0	113.0	220	2	O	536.36
165	DBN	105.0	101.0	139	1	O	755.40
166	DBN	130.0	120.0	142	2	O	915.49
167	DBN	120.0	115.0	231	3	O	519.48
168	DBN	110.0	93.0	153	3	O	718.95
169	DBN	212.0	210.0	221	2	N	959.28
170	DBN	150.0	129.0	201	3	O	746.27
171	DBN	108.5	117.0	146	2	O	743.15
172	DBN	260.0	235.0	305	3	O	852.46
173	DBN	135.0	120.0	123	2	O	1 097.56
174	DBN	180.0	175.0	311	3	N	578.78
175	DBN	154.0	130.5	109	1	N	1 412.84
176	DBN	117.0	112.0	121	3	O	966.94
177	DBN	150.0	145.0	201	3	O	746.27
178	DBN	193.0	187.0	301	2	O	641.20
179	DBN	85.0	145.0	125	3	O	680.00
180	DBN	100.0	98.0	98	2	N	1 020.41
181	DBN	170.0	160.0	210	3	O	809.52
182	DBN	163.0	133.0	188	2	O	867.02
183	DBN	170.0	145.0	195	2	O	871.79
184	DBN	177.7	172.0	264	2	N	673.11
185	DBN	170.0	145.0	187	3	O	909.09
186	DBN	290.0	260.0	266	3	O	1 090.23
187	DBN	255.0	245.0	307	2	O	830.62
188	DBN	189.0	184.0	240	2	O	787.50
189	DBN	135.0	132.0	167	2	O	808.38
190	DBN	155.0	145.0	155	2	O	1 000.00
191	DBN	110.0	106.0	158	2	O	696.20
192	DBN	112.5	85.0	144	1	O	781.25
193	DBN	350.0	205.0	241	3	O	1 452.28
194	DBN	350.0	123.0	179	2	O	1 955.31
195	DBN	95.0	85.0	200	2	O	475.00
196	DBN	205.0	180.0	217	2	O	944.70
197	DBN	105.0	90.0	101	2	O	1 039.60
198	DBN	80.0	80.0	99	3	N	808.08
199	DBN	175.0	148.0	238	2	O	735.29
200	DBN	68.0	154.0	162	2	O	419.75

No	City	Price	Value	Area	Size	Age	PSM
201	DBN	260.0	240.0	253	2	O	1 027.67
202	DBN	245.0	230.0	216	2	O	1 134.26
203	DBN	80.0	77.0	119	1	N	672.27
204	DBN	97.0	93.0	173	2	O	560.69
205	DBN	130.0	125.0	175	2	O	742.86
206	DBN	77.5	74.5	106	1	O	731.13
207	DBN	110.0	100.0	167	2	O	658.68
208	DBN	221.6	216.6	271	2	O	817.71
209	DBN	94.0	89.0	185	2	O	508.11
210	DBN	98.0	97.0	142	1	O	690.14
211	DBN	226.0	206.0	311	2	O	726.69
212	DBN	200.0	185.0	213	2	O	938.97
213	DBN	99.0	120.0	170	2	O	582.35
214	DBN	85.0	78.0	98	1	O	867.35
215	DBN	120.0	115.0	288	1	N	416.67
216	DBN	92.5	75.0	114	2	O	811.40
217	DBN	165.0	155.0	149	2	O	1 107.38
218	DBN	340.0	330.0	334	3	O	1 017.96
219	DBN	95.0	93.0	118	1	O	805.08
220	DBN	70.0	67.0	95	1	N	736.84
221	DBN	145.0	127.0	136	2	O	1 066.18
222	DBN	75.0	73.0	114	1	N	657.89
223	DBN	68.5	68.5	81	1	O	845.68
224	DBN	215.0	215.0	250	2	N	860.00
225	DBN	230.0	279.0	200	2	O	1 150.00
226	DBN	275.0	215.0	255	3	O	1 078.43
227	DBN	120.0	105.0	142	1	O	845.07
228	DBN	235.0	224.0	303	2	N	775.58
229	DBN	132.0	112.0	179	1	O	737.43
230	DBN	118.0	108.0	120	1	O	983.33
231	DBN	120.0	117.0	130	2	O	923.08
232	DBN	119.5	119.5	90	1	O	1 327.78
233	DBN	198.0	150.0	203	3	O	975.37
234	DBN	125.0	103.0	138	2	O	905.80
235	DBN	260.0	232.0	237	3	O	1 097.05
236	DBN	218.0	207.0	195	1	O	1 117.95
237	DBN	80.0	75.0	149	1	N	536.91
238	DBN	100.0	96.0	166	1	O	602.41
239	DBN	58.0	57.0	104	1	O	557.69
240	DBN	120.0	115.0	192	3	O	625.00
241	DBN	120.0	120.0	123	2	N	975.61
242	DBN	107.0	97.0	95	2	O	1 126.32
243	DBN	326.4	250.0	168	1	N	1 942.86
244	DBN	94.0	93.0	119	2	O	789.92
245	DBN	88.0	131.0	178	3	O	494.38
246	DBN	120.0	117.0	98	3	N	1 224.49
247	DBN	150.0	119.0	209	2	O	717.70
248	DBN	155.0	127.0	222	2	O	698.20
249	DBN	130.0	117.0	171	2	O	760.23
250	DBN	123.0	113.0	141	3	O	872.34

No	City	Price	Value	Area	Size	Age	PSM
251	DBN	200.0	195.0	214	3	O	934.58
252	DBN	135.0	110.0	156	3	O	865.38
253	DBN	105.0	103.0	130	2	O	807.69
254	DBN	115.0	90.0	110	3	O	1 045.45
255	DBN	165.0	145.0	194	3	O	850.52
256	DBN	190.0	185.0	153	2	O	1 241.83
257	DBN	180.0	175.0	190	3	O	947.37
258	DBN	135.0	98.0	169	2	O	798.82
259	DBN	258.0	221.0	175	1	O	1 474.29
260	DBN	315.0	290.0	269	2	O	1 171.00
261	DBN	117.0	150.0	216	2	O	541.67
262	DBN	205.0	186.0	176	1	O	1 164.77
263	DBN	130.0	125.0	175	2	O	742.86
264	DBN	100.0	95.0	159	1	N	628.93
265	DBN	205.0	195.0	132	2	O	1 553.03
266	DBN	105.0	95.0	82	1	O	1 280.49
267	DBN	98.0	98.0	138	3	O	710.14
268	DBN	165.0	143.0	149	2	O	1 107.38
269	DBN	65 0	65.0	88	1	O	738.64
270	DBN	150.0	127.0	199	2	O	753.77
271	DBN	395.0	385.0	212	3	O	1 863.21
272	DBN	88.3	85.0	108	1	O	817.59
273	DBN	205.0	200.0	190	3	N	1 078.95
274	DBN	65.0	55.0	94	1	O	691.49
275	DBN	150.0	147.0	126	1	O	1 190.48
276	DBN	94.0	85.0	114	1	O	824.56
277	DBN	79.0	73.0	92	1	O	858.70
278	DBN	126.0	121.0	151	2	O	834.44
279	DBN	155.0	142.0	162	2	O	956.79
280	DBN	110.0	105.0	228	2	O	482.46
281	DBN	140.0	115.0	157	2	O	891.72
282	DBN	128.0	126.0	164	1	N	780.49
283	DBN	110.0	104.0	118	3	O	932.20
284	DBN	130.0	125.0	158	3	O	822.78
285	DBN	137.5	117.5	127	2	O	1 082.68
286	DBN	90.0	85.0	127	1	O	708.66
287	DBN	117.0	109.0	197	1	O	593.91
288	DBN	100.0	93.0	220	2	O	454.55
289	DBN	136.0	136.0	145	2	N	937.93
290	DBN	62.0	65.4	80	1	O	775.00
291	DBN	71.9	71.9	90	1	N	798.89
292	DBN	126.0	119.0	154	2	O	818.18
293	DBN	90.0	88.0	183	1	O	491.80
294	DBN	100.0	100.0	142	2	O	704.23
295	DBN	143.0	126.0	215	3	O	665.12
296	DBN	160.0	142.0	320	2	O	500.00
297	DBN	120.0	80.0	81	2	O	1 481.48
298	DBN	65.0	63.0	85	1	O	764.71
299	DBN	105.0	99.0	176	1	O	596.59
300	DBN	64.0	56.0	92	1	O	695.65

No	City	Price	Value	Area	Size	Age	PSM
301	DBN	138.0	131.0	138	1	O	1 000.00
302	DBN	140.0	130.0	246	2	O	569.11
303	DBN	183.0	158.0	198	3	O	924.24
304	DBN	110.0	103.0	149	1	O	738.26
305	DBN	172.0	125.0	160	2	O	1 075.00
306	DBN	135.0	112.0	169	2	O	798.82
307	DBN	130.0	120.0	258	3	O	503.88
308	DBN	129.0	124.0	146	3	O	883.56
309	DBN	182.0	157.0	201	2	O	905.47
310	DBN	141.9	141.5	121	3	N	1 172.73
311	DBN	152.0	127.0	165	2	O	921.21
312	DBN	70.0	90.0	112	2	O	625.00
313	DBN	112.0	107.0	139	2	O	805.76
314	DBN	105.0	99.0	159	2	O	660.38
315	DBN	180.0	155.0	208	2	O	865.38
316	DBN	187.0	162.0	213	2	O	877.93
317	DBN	95.0	90.0	116	2	O	818.97
318	DBN	191.5	166.0	140	3	O	1 367.86
319	DBN	205.0	200.0	264	3	O	776.52
320	DBN	125.0	115.0	143	2	O	874.13
321	DBN	145.0	128.0	260	2	O	557.69
322	DBN	135.0	115.0	182	3	O	741.76
323	DBN	112.0	100.0	137	2	O	817.52
324	DBN	106.0	100.0	88	2	N	1 204.55
325	DBN	110.2	110.0	80	3	N	1 377.50
326	DBN	257.0	224.0	183	3	O	1 404.37
327	DBN	135.0	115.0	169	3	O	798.82
328	DBN	267.5	257.5	256	3	O	1 044.92
329	DBN	110.0	108.0	153	3	O	718.95
330	DBN	143.0	133.0	159	2	O	899.37
331	DBN	220.0	200.0	215	3	O	1 023.26
332	DBN	125.0	140.0	223	3	O	560.54
333	DBN	124.0	104.0	162	2	O	765.43
334	DBN	110.0	98.0	157	3	O	700.64
335	DBN	140.0	131.0	246	2	O	569.11
336	DBN	130.0	125.0	132	1	O	984.85
337	DBN	175.0	162.0	274	2	O	638.69
338	DBN	165.0	145.0	202	2	O	816.83
339	DBN	195.2	194.0	297	2	N	657.24
340	DBN	135.0	117.0	134	2	O	1 007.46
341	DBN	130.0	125.0	157	2	O	828.03
342	DBN	110.0	104.0	143	1	O	769.23
343	DBN	140.0	128.0	224	2	O	625.00
344	DBN	106.0	106.0	83	2	N	1 277.11
345	DBN	89.0	145.0	215	2	O	413.95
346	DBN	100.0	97.0	185	2	O	540.54
347	DBN	96.0	83.0	97	3	O	989.69
348	DBN	153.0	143.0	196	2	O	780.61
349	DBN	140.0	125.0	150	1	O	933.33
350	DBN	155.0	148.0	142	3	O	1 091.55

No	City	Price	Value	Area	Size	Age	PSM
351	DBN	210.0	185.0	185	2	O	1 135.14
352	DBN	90.0	125.0	137	3	O	656.93
353	DBN	133.5	128.5	153	1	O	872.55
354	DBN	80.0	76.0	87	1	O	919.54
355	DBN	162.4	185.0	220	2	O	738.18
356	DBN	155.0	150.0	144	1	O	1 076.39
357	DBN	100.0	95.0	154	1	N	649.35
358	DBN	85.0	80.0	88	1	O	965.91
359	DBN	155.0	340.0	278	3	O	557.55
360	DBN	150.0	215.0	181	2	O	828.73
361	DBN	65.0	65.0	87	1	O	747.13
362	DBN	230.0	185.0	188	1	O	1 223.40
363	DBN	300.0	280.0	290	3	O	1 034.48
364	DBN	85.0	82.0	94	1	O	904.26
365	DBN	220.0	250.0	258	2	O	852.71
366	DBN	90.0	80.0	122	1	N	737.70
367	DBN	88.0	80.0	133	1	O	661.65
368	DBN	100.0	96.0	166	1	O	602.41
369	DBN	110.0	190.0	306	3	O	359.48
370	DBN	65.0	65.0	90	2	N	722.22
371	DBN	172.0	162.0	277	3	O	620.94
372	DBN	93.0	71.0	90	1	O	1 033.33
373	DBN	195.0	148.0	205	2	O	951.22
374	DBN	170.0	150.0	170	1	O	1 000.00
375	DBN	110.0	105.0	228	2	O	482.46
376	DBN	130.0	105.0	102	2	N	1 274.51

No	City	Price	Value	Area	Size	Age	PSM
1	JHB	195.0	162.0	213	3	O	915.49
2	JHB	145.0	140.0	162	1	O	895.06
3	JHB	100.0	95.0	177	1	O	564.97
4	JHB	250.0	190.0	203	1	O	1 231.53
5	JHB	200.0	187.0	258	2	O	775.19
6	JHB	290.0	272.0	238	3	O	1 218.49
7	JHB	110.0	187.0	145	1	O	758.62
8	JHB	72.0	66.0	85	1	O	847.06
9	JHB	190.0	179.0	175	1	O	1 085.71
10	JHB	50.0	50.0	88	1	O	568.18
11	JHB	190.0	182.0	178	2	O	1 067.42
12	JHB	385.0	360.0	241	3	O	1 597.51
13	JHB	90.0	80.0	147	3	O	612.24
14	JHB	214.0	208.0	232	2	O	922.41
15	JHB	225.0	220.0	278	2	O	809.35
16	JHB	212.0	193.0	216	2	O	981.48
17	JHB	212.0	193.0	216	2	O	981.48
18	JHB	118.0	105.0	156	1	O	756.41
19	JHB	115.0	100.0	160	2	O	718.75
20	JHB	65.0	138.0	194	1	O	335.05
21	JHB	97.9	93.0	123	1	O	795.93
22	JHB	110.0	131.0	202	2	O	544.55
23	JHB	95.0	93.0	177	1	O	536.72
24	JHB	132.0	132.0	144	1	O	916.67
25	JHB	128.0	118.0	157	1	O	815.29
26	JHB	280.0	240.0	295	3	O	949.15
27	JHB	99.0	143.0	178	2	O	556.18
28	JHB	180.0	415.0	261	3	O	689.66
29	JHB	280.0	260.0	247	2	O	1 133.60
30	JHB	200.0	189.0	200	3	O	1 000.00
31	JHB	200.0	190.0	186	2	O	1 075.27
32	JHB	220.0	214.0	250	2	O	880.00
33	JHB	330.0	303.0	227	3	O	1 453.74
34	JHB	245.0	165.0	172	2	O	1 424.42
35	JHB	240.0	210.0	212	2	O	1 132.08
36	JHB	277.0	224.0	172	1	O	1 610.47
37	JHB	147.5	127.0	171	1	O	862.57
38	JHB	260.0	238.0	328	3	O	792.68
39	JHB	212.0	188.0	206	1	O	1 029.13
40	JHB	255.0	233.0	234	X	N	1 089.74
41	JHB	230.0	213.0	202	X	O	1 138.61
42	JHB	335.0	300.0	287	4	N	1 167.25
43	JHB	183.0	145.0	154	3	O	1 188.31
44	JHB	118.0	150.0	172	X	O	686.05
45	JHB	208.5	201.5	149	1	O	1 399.33
46	JHB	250.0	238.0	308	3	O	811.69
47	JHB	75.5	64.0	149	1	O	506.71
48	JHB	80.0	77.0	116	1	O	689.66
49	JHB	114.0	111.0	117	X	N	974.36
50	JHB	100.0	96.0	196	1	O	510.20

No	City	Price	Value	Area	Size	Age	PSM
51	JHB	63.0	94.0	192	1	O	328.12
52	JHB	123.0	114.0	258	1	O	476.74
53	JHB	160.0	115.0	320	2	O	500.00
54	JHB	100.0	94.0	184	1	O	543.48
55	JHB	82.0	74.0	100	1	O	820.00
56	JHB	95.0	92.0	166	1	O	572.29
57	JHB	165.5	147.0	267	2	O	619.85
58	JHB	133.0	129.0	234	2	O	568.38
59	JHB	110.0	104.0	220	1	O	500.00
60	JHB	46.0	73.5	158	2	O	291.14
61	JHB	125.0	112.0	151	2	O	827.81
62	JHB	100.0	97.0	178	2	O	561.80
63	JHB	165.0	136.0	185	2	O	891.89
64	JHB	105.0	92.0	197	1	O	532.99
65	JHB	64.0	75.0	181	2	O	353.59
66	JHB	85.0	76.0	148	1	O	574.32
67	JHB	83.0	62.0	158	1	O	525.32
68	JHB	110.0	101.5	167	1	O	658.68
69	JHB	141.0	128.0	266	1	O	530.08
70	JHB	70.0	66.0	220	1	O	318.18
71	JHB	332.0	330.0	229	1	N	1 449.78
72	JHB	232.0	228.0	235	2	N	987.23
73	JHB	255.0	233.0	255	1	O	1 000.00
74	JHB	200.0	195.0	293	2	O	682.59
75	JHB	67.0	112.0	210	1	O	319.05
76	JHB	155.0	148.0	255	1	O	607.84
77	JHB	250.0	242.0	209	2	O	1 196.17
78	JHB	140.0	127.0	166	1	O	843.37
79	JHB	290.0	243.0	264	2	O	1 098.48
80	JHB	110.0	95.0	209	1	O	526.32
81	JHB	140.0	134.0	212	1	O	660.38
82	JHB	165.0	140.0	249	1	O	662.65
83	JHB	105.0	101.0	210	1	O	500.00
84	JHB	155.0	149.0	185	1	O	837.84
85	JHB	173.0	170.0	209	3	O	827.75
86	JHB	115.0	98.0	131	2	O	877.86
87	JHB	84.7	81.5	131	1	O	646.56
88	JHB	120.0	111.0	149	2	O	805.37
89	JHB	135.0	112.0	135	1	O	1 000.00
90	JHB	110.0	105.0	231	2	O	476.19
91	JHB	160.0	150.0	256	1	O	625.00
92	JHB	150.0	150.0	135	2	N	1 111.11
93	JHB	160.0	144.0	107	3	O	1 495.33
94	JHB	171.0	151.0	165	2	O	1 036.36
95	JHB	106.0	88.0	82	2	O	1 292.68
96	JHB	206.3	203.0	168	1	N	1 227.98
97	JHB	198.0	193.0	140	1	N	1 414.29
98	JHB	260.0	232.0	202	3	O	1 287.13
99	JHB	332.0	312.0	247	2	O	1 344.13
100	JHB	162.1	158.1	123	2	N	1 317.89

No	City	Price	Value	Area	Size	Age	PSM
101	JHB	415.0	325.0	238	2	O	1 743.70
102	JHB	141.7	134.7	80	2	N	1 771.25
103	JHB	235.0	231.0	176	2	N	1 335.23
104	JHB	160.0	152.0	128	1	O	1 250.00
105	JHB	325.0	197.0	219	2	O	1 484.02
106	JHB	178.0	174.0	138	2	N	1 289.86
107	JHB	135.6	128.6	84	2	N	1 614.29
108	JHB	140.0	158.0	219	2	O	639.27
109	JHB	130.0	122.0	117	2	O	1 111.11
110	JHB	270.0	224.0	180	2	O	1 500.00
111	JHB	114.5	107.5	82	2	N	1 396.34
112	JHB	138.0	135.0	109	2	N	1 266.06
113	JHB	396.0	366.0	251	4	N	1 577.69
114	JHB	226.0	226.0	132	3	N	1 712.12
115	JHB	190.0	161.0	163	2	O	1 165.64
116	JHB	137.5	124.5	119	2	O	1 155.46
117	JHB	185.0	153.0	165	2	O	1 121.21
118	JHB	135.0	126.0	138	2	O	978.26
119	JHB	130.0	127.0	109	1	O	1 192.66
120	JHB	230.0	207.0	290	3	O	793.10
121	JHB	260.0	228.0	301	3	O	863.79
122	JHB	195.0	163.0	203	2	O	960.59
123	JHB	165.0	157.0	158	2	O	1 044.30
124	JHB	142.0	127.0	120	2	O	1 183.33
125	JHB	335.0	263.0	303	2	O	1 105.61
126	JHB	150.0	113.5	101	2	O	1 485.15
127	JHB	133.0	110.0	96	2	O	1 385.42
128	JHB	133.0	126.0	146	2	O	910.96
129	JHB	33.5	200.0	183	3	O	183.06
130	JHB	170.0	161.5	190	3	O	894.74
131	JHB	175.0	196.0	238	3	O	735.29
132	JHB	76.0	74.0	85	1	O	894.12
133	JHB	237.0	199.0	256	2	O	925.78
134	JHB	30.6	180.0	112	2	O	273.21
135	JHB	160.5	155.0	122	2	N	1 315.57
136	JHB	150.0	48.0	203	1	O	738.92
137	JHB	75.0	70.0	80	1	O	937.50
138	JHB	120.0	118.0	200	1	O	600.00
139	JHB	65.0	64.0	168	1	O	386.90
140	JHB	185.0	152.0	147	3	O	1 258.50
141	JHB	105.0	95.0	133	1	O	789.47
142	JHB	107.2	98.0	81	1	O	1 323.46
143	JHB	158.0	138.0	159	1	O	993.71
144	JHB	447.5	417.5	225	2	O	1 988.89
145	JHB	45.0	155.0	265	1	O	169.81
146	JHB	435.0	398.0	263	2	O	1 653.99
147	JHB	88.0	78.0	146	1	O	602.74
148	JHB	156.0	147.0	123	1	N	1 268.29
149	JHB	85.0	103.0	144	1	O	590.28
150	JHB	225.0	218.0	139	1	O	1 618.71

No	City	Price	Value	Area	Size	Age	PSM
151	JHB	240.0	235.0	188	1	N	1 276.60
152	JHB	210.0	210.0	156	1	N	1 346.15
153	JHB	240.0	235.0	188	1	N	1 276.60
154	JHB	240.0	235.0	188	1	N	1 276.60
155	JHB	118.0	130.0	130	1	O	907.69
156	JHB	205.0	175.0	208	1	O	985.58
157	JHB	119.0	108.8	101	1	N	1 178.22
158	JHB	87.0	83.0	229	1	O	379.91
159	JHB	155.0	140.0	170	1	O	911.76
160	JHB	170.0	162.0	183	1	O	928.96
161	JHB	210.0	207.0	217	2	O	967.74
162	JHB	74.0	58.0	90	2	O	822.22
163	JHB	92.0	85.0	85	1	O	1 082.35
164	JHB	58.0	56.5	80	1	O	725.00
165	JHB	125.0	118.2	222	2	O	563.06
166	JHB	52.5	49.5	99	1	O	530.30
167	JHB	163.0	129.0	170	2	O	958.82
168	JHB	160.0	150.0	240	2	O	666.67
169	JHB	96.0	121.0	166	2	O	578.31
170	JHB	135.0	131.0	262	1	O	515.27
171	JHB	318.0	285.0	320	2	O	993.75
172	JHB	148.0	137.0	156	2	O	948.72
173	JHB	69.0	90.0	147	2	O	469.39
174	JHB	155.0	155.0	183	2	O	846.99
175	JHB	150.0	217.0	255	2	O	588.24
176	JHB	117.0	144.0	201	3	O	582.09
177	JHB	75.0	60.5	107	1	O	700.93
178	JHB	109.0	104.0	94	2	O	1 159.57
179	JHB	58.0	177.0	242	2	O	239.67
180	JHB	122.0	116.0	93	2	O	1 311.83
181	JHB	315.0	235.0	293	3	O	1 075.09
182	JHB	147.5	141.0	103	1	N	1 432.04
183	JHB	340.0	313.0	247	3	O	1 376.52
184	JHB	82.0	138.0	143	2	O	573.43
185	JHB	210.0	176.0	193	2	O	1 088.08
186	JHB	159.0	315.0	287	2	O	554.01
187	JHB	167.0	220.0	229	2	O	729.26
188	JHB	114.0	132.0	207	2	O	550.72
189	JHB	235.0	215.0	212	1	O	1 108.49
190	JHB	255.0	233.0	134	2	N	1 902.99
191	JHB	30.0	225.0	217	3	O	138.25
192	JHB	240.0	219.0	158	2	O	1 518.99
193	JHB	390.0	347.0	188	1	N	2 074.47
194	JHB	220.0	410.0	206	2	O	1 067.96
195	JHB	180.0	163.0	162	1	O	1 111.11
196	JHB	256.0	246.0	266	2	O	962.41
197	JHB	90.0	82.0	195	1	O	461.54
198	JHB	150.0	137.0	207	2	O	724.64
199	JHB	115.0	102.0	103	2	O	1 116.50
200	JHB	131.0	123.0	243	1	O	539.09

No	City	Price	Value	Area	Size	Age	PSM
201	JHB	90.0	86.0	86	3	O	1 046.51
202	JHB	115.0	106.0	267	1	O	430.71
203	JHB	76.0	72.0	121	1	O	628.10
204	JHB	125.0	164.0	301	2	O	415.28
205	JHB	144.5	138.5	150	2	O	963.33
206	JHB	68.0	61.0	100	2	O	680.00
207	JHB	150.0	126.5	160	2	O	937.50
208	JHB	53.0	243.0	329	2	O	161.09
209	JHB	235.0	206.0	249	2	O	943.78
210	JHB	61.0	93.0	187	1	O	326.20
211	JHB	87.0	81.0	110	1	O	790.91
212	JHB	68.5	65.0	119	1	O	575.63
213	JHB	102.0	87.0	132	1	O	772.73
214	JHB	77.0	71.0	161	1	O	478.26
215	JHB	135.0	127.0	145	2	O	931.03
216	JHB	85.0	72.0	112	1	O	758.93
217	JHB	258.0	245.0	227	2	N	1 136.56
218	JHB	100.0	95.0	188	1	O	531.91
219	JHB	185.0	177.0	200	2	O	925.00
220	JHB	140.0	125.0	186	2	O	752.69
221	JHB	87.0	69.0	92	1	O	945.65
222	JHB	76.0	71.0	109	1	O	697.25
223	JHB	135.0	118.0	142	1	O	950.70
224	JHB	212.0	163.0	199	2	O	1 065.33
225	JHB	85.0	78.0	159	1	O	534.59
226	JHB	59.0	124.0	167	2	O	353.29
227	JHB	193.0	172.0	233	2	O	828.33
228	JHB	85.0	82.0	208	1	O	408.65
229	JHB	170.0	127.0	164	2	O	1 036.59
230	JHB	160.0	135.0	209	2	O	765.55
231	JHB	438.0	384.0	264	3	O	1 659.09
232	JHB	60.0	58.0	225	1	O	266.67
233	JHB	225.0	192.0	225	2	O	1 000.00
234	JHB	125.0	122.0	192	1	O	651.04
235	JHB	142.5	137.0	213	1	O	669.01
236	JHB	215.0	185.0	230	2	O	934.78
237	JHB	108.0	102.0	161	1	O	670.81
238	JHB	190.0	162.0	332	2	O	572.29
239	JHB	176.0	156.0	203	1	O	867.00
240	JHB	170.0	165.0	193	1	O	880.83
241	JHB	150.0	141.0	189	1	O	793.65
242	JHB	131.0	124.0	165	2	O	793.94
243	JHB	175.0	158.0	298	2	O	587.25
244	JHB	124.0	122.0	162	1	O	765.43
245	JHB	245.0	218.0	267	1	O	917.60
246	JHB	240.0	213.0	260	2	O	923.08
247	JHB	145.0	139.0	206	1	O	703.88
248	JHB	171.5	162.0	172	1	O	997.09
249	JHB	129.0	97.0	159	2	O	754.72
250	JHB	120.0	108.0	131	2	O	916.03

No	City	Price	Value	Area	Size	Age	PSM
251	JHB	190.0	179.0	213	1	O	892.02
252	JHB	125.0	116.0	178	2	O	702.25
253	JHB	190.0	166.0	181	1	O	1 049.72
254	JHB	180.0	166.0	257	2	O	700.39
255	JHB	145.0	110.0	143	1	O	1 013.99
256	JHB	375.0	338.0	293	3	O	1 279.86
257	JHB	166.0	166.0	136	2	N	1 220.59
258	JHB	230.0	215.0	313	2	O	734.82
259	JHB	215.0	195.0	177	2	O	1 214.69
260	JHB	190.0	186.0	117	2	N	1 623.93
261	JHB	108.0	97.0	85	2	O	1 270.59
262	JHB	380.0	365.0	270	3	O	1 407.41
263	JHB	145.0	141.0	105	1	O	1 380.95
264	JHB	182.6	179.0	116	2	N	1 574.14
265	JHB	200.0	190.0	189	2	O	1 058.20
266	JHB	455.0	455.0	300	2	N	1 516.67
267	JHB	199.7	196.0	176	2	N	1 134.66
268	JHB	162.5	156.5	168	2	O	967.26
269	JHB	220.0	213.0	255	2	O	862.75
270	JHB	175.0	165.0	216	3	O	810.19
271	JHB	163.7	159.7	111	2	N	1 474.77
272	JHB	120.6	113.6	80	2	N	1 507.50
273	JHB	158.0	114.0	95	2	O	1 663.16
274	JHB	139.3	135.3	104	2	N	1 339.42
275	JHB	172.6	169.0	128	2	N	1 348.44
276	JHB	100.0	138.0	99	2	O	1 010.10
277	JHB	191.0	187.0	149	2	N	1 281.88
278	JHB	131.1	124.1	80	2	N	1 638.75
279	JHB	186.5	182.5	161	2	N	1 158.39
280	JHB	220.0	196.0	201	2	O	1 094.53
281	JHB	360.0	330.0	289	3	O	1 245.67
282	JHB	127.0	118.0	185	2	O	686.49
283	JHB	135.0	126.0	128	2	O	1 132.81
284	JHB	165.0	145.0	171	1	O	906.43
285	JHB	78.0	72.0	110	1	O	690.91
286	JHB	97.0	153.0	197	2	O	492.39
287	JHB	259.0	223.0	262	2	O	954.20
288	JHB	100.0	226.0	214	1	O	467.29
289	JHB	125.0	115.0	131	2	O	954.20
290	JHB	99.5	95.5	150	1	O	663.33
291	JHB	90.0	75.0	110	1	O	818.18
292	JHB	95.0	71.0	130	1	O	730.77
293	JHB	189.0	155.0	184	1	O	1 027.17
294	JHB	95.0	90.0	195	1	O	487.18
295	JHB	98.1	98.1	85	1	N	1 154.12
296	JHB	385.0	274.0	276	3	O	1 105.07
297	JHB	75.0	72.0	143	1	O	524.48
298	JHB	245.0	232.0	243	3	N	1 008.23
299	JHB	190.0	176.0	216	2	O	879.63
300	JHB	135.0	120.0	213	1	O	633.80

No	City	Price	Value	Area	Size	Age	PSM
301	JHB	240.0	220.0	331	2	O	725.08
302	JHB	89.0	85.0	99	2	O	898.99
303	JHB	85.0	80.0	130	1	O	653.85
304	JHB	70.0	102.0	152	1	O	460.53
305	JHB	80.0	77.0	131	1	O	610.69
306	JHB	110.0	100.0	145	2	O	758.62
307	JHB	165.0	150.0	106	1	O	1 556.60
308	JHB	105.0	90.0	138	1	O	760.87
309	JHB	135.0	119.0	118	1	O	1 144.07
310	JHB	96.8	95.0	85	1	N	1 138.82
311	JHB	100.0	91.0	100	1	O	1 000.00
312	JHB	103.0	99.0	138	1	O	746.38
313	JHB	275.0	245.0	210	2	O	1 309.52
314	JHB	423.0	317.0	258	2	O	1 639.53
315	JHB	131.0	126.0	133	1	O	984.96
316	JHB	360.0	315.0	242	2	O	1 487.60
317	JHB	230.0	215.0	180	2	O	1 277.78
318	JHB	115.0	107.0	94	1	O	1 223.40
319	JHB	152.0	141.0	113	2	O	1 345.13
320	JHB	152.0	141.0	123	2	O	1 235.77
321	JHB	118.0	111.0	119	1	O	991.60
322	JHB	85.0	150.0	170	2	O	500.00
323	JHB	125.0	128.0	164	2	O	762.20
324	JHB	192.0	175.0	231	2	O	831.17
325	JHB	290.0	250.0	239	2	O	1 213.39
326	JHB	110.0	104.0	143	1	N	769.23
327	JHB	154.0	139.0	172	2	O	895.35
328	JHB	170.0	142.0	164	2	O	1 036.59
329	JHB	95.0	88.0	146	2	O	650.68
330	JHB	225.0	185.0	209	2	O	1 076.56
331	JHB	75.0	70.0	89	1	O	842.70
332	JHB	114.7	114.0	135	2	N	849.63
333	JHB	95.0	92.0	132	2	O	719.70
334	JHB	300.0	249.0	239	2	O	1 255.23
335	JHB	152.0	147.0	120	2	N	1 266.67
336	JHB	85.0	151.0	238	2	O	357.14
337	JHB	236.0	183.0	207	2	O	1 140.10
338	JHB	154.0	154.0	182	1	N	846.15
339	JHB	81.0	115.0	104	2	O	778.85
340	JHB	120.0	117.0	217	1	O	553.00
341	JHB	185.0	213.0	184	1	O	1 005.43
342	JHB	265.0	239.0	276	1	N	960.14
343	JHB	155.0	130.0	143	2	O	1 083.92
344	JHB	280.0	249.0	219	2	O	1 278.54
345	JHB	227.0	207.0	294	2	O	772.11
346	JHB	78.0	68.0	148	1	O	527.03
347	JHB	330.0	300.0	312	2	O	1 057.69
348	JHB	185.0	170.0	290	1	O	637.93
349	JHB	255.0	245.0	218	2	O	1 169.72
350	JHB	165.0	205.0	193	2	O	854.92

No	City	Price	Value	Area	Size	Age	PSM
351	JHB	400.0	362.0	325	2	O	1 230.77
352	JHB	130.0	121.0	95	2	C	1 368.42
353	JHB	190.0	185.0	273	1	O	695.97
354	JHB	280.0	213.0	283	2	O	989.40
355	JHB	238.0	213.0	159	1	O	1 496.86
356	JHB	190.0	178.0	205	2	O	926.83
357	JHB	370.0	306.5	277	3	O	1 335.74
358	JHB	150.0	120.0	226	2	O	663.72
359	JHB	300.0	280.0	253	2	O	1 185.77
360	JHB	210.0	183.0	213	2	O	985.92
361	JHB	113.0	92.0	99	2	O	1 141.41
362	JHB	214.0	190.0	208	2	O	1 028.85
363	JHB	112.0	108.0	216	1	O	518.52
364	JHB	130.0	114.0	314	1	O	414.01
365	JHB	107.5	81.0	80	1	O	1 343.75
366	JHB	95.0	81.0	151	1	O	629.14
367	JHB	115.0	96.0	242	2	O	475.21
368	JHB	60.0	110.0	165	2	O	363.64
369	JHB	73.4	71.0	125	1	O	587.20
370	JHB	202.0	179.0	244	2	O	827.87
371	JHB	218.0	183.0	252	2	O	865.08
372	JHB	160.0	131.0	185	2	O	864.86
373	JHB	149.0	125.0	244	2	O	610.66
374	JHB	112.0	97.0	166	1	O	674.70
375	JHB	260.0	247.0	290	2	O	896.55
376	JHB	105.0	96.0	211	1	O	497.63
377	JHB	65.0	136.5	179	1	O	363.13
378	JHB	95.0	92.0	170	2	O	558.82
379	JHB	145.0	121.0	198	1	O	732.32
380	JHB	425.0	378.0	343	3	O	1 239.07
381	JHB	135.0	128.0	234	1	O	576.92
382	JHB	398.0	402.0	299	3	O	1 331.10
383	JHB	110.0	108.0	217	1	O	506.91
384	JHB	126.0	121.0	165	1	O	763.64
385	JHB	140.0	137.0	214	1	O	654.21
386	JHB	480.0	445.0	349	3	O	1 375.36
387	JHB	92.5	88.0	137	1	O	675.18
388	JHB	195.0	175.0	275	3	O	709.09
389	JHB	160.0	153.0	171	1	O	935.67
390	JHB	205.0	182.0	297	2	O	690.24
391	JHB	130.0	115.0	200	2	O	650.00
392	JHB	202.5	192.0	228	1	O	888.16
393	JHB	98.0	85.0	91	3	O	1 076.92
394	JHB	145.0	135.0	169	1	O	857.99
395	JHB	190.0	181.0	228	1	O	833.33
396	JHB	200.0	192.0	211	1	O	947.87
397	JHB	195.0	168.0	223	2	O	874.44
398	JHB	263.2	259.2	200	2	N	1 316.00
399	JHB	350.0	324.0	332	3	O	1 054.22
400	JHB	125.0	106.0	124	2	O	1 008.06

No	City	Price	Value	Area	Size	Age	PSM
401	JHB	268.0	257.0	198	2	O	1 353.54
402	JHB	178.0	174.0	121	2	N	1 471.07
403	JHB	255.0	232.0	158	1	O	1 613.92
404	JHB	165.0	160.0	243	3	O	679.01
405	JHB	310.0	305.0	338	1	N	917.16
406	JHB	156.7	141.0	147	1	O	1 065.99
407	JHB	345.0	315.0	317	3	N	1 088.33
408	JHB	240.0	221.5	217	2	O	1 105.99
409	JHB	217.5	183.0	222	2	O	979.73
410	JHB	500.0	420.0	342	3	O	1 461.99
411	JHB	135.0	133.0	202	2	O	668.32
412	JHB	215.0	175.0	181	2	O	1 187.85
413	JHB	110.0	102.0	170	1	O	647.06
414	JHB	220.0	201.0	242	2	O	909.09
415	JHB	186.0	158.0	180	2	O	1 033.33
416	JHB	100.0	96.0	166	2	O	602.41
417	JHB	243.6	238.0	175	2	N	1 392.00
418	JHB	205.0	190.0	223	3	O	919.28
419	JHB	270.0	246.0	231	3	O	1 168.83
420	JHB	250.0	200.0	237	2	O	1 054.85
421	JHB	150.0	135.0	233	1	O	643.78
422	JHB	181.0	179.0	118	2	N	1 533.90
423	JHB	130.0	118.0	106	2	O	1 226.42
424	JHB	170.0	143.0	164	2	O	1 036.59
425	JHB	440.0	327.0	290	3	O	1 517.24
426	JHB	279.0	279.0	250	2	N	1 116.00
427	JHB	285.0	255.0	260	3	O	1 096.15
428	JHB	345.0	288.0	296	3	O	1 165.54
429	JHB	279.0	249.0	220	3	O	1 268.18
430	JHB	157.0	129.0	130	2	O	1 207.69
431	JHB	105.0	102.0	189	1	O	555.56
432	JHB	195.0	180.0	209	1	O	933.01
433	JHB	113.0	105.0	162	1	O	697.53
434	JHB	90.0	83.0	135	1	O	666.67
435	JHB	50.0	113.0	168	1	O	297.62
436	JHB	65.0	87.0	167	1	O	389.22
437	JHB	270.0	225.0	178	2	O	1 516.85
438	JHB	100.0	95.0	172	1	O	581.40
439	JHB	95.0	83.0	94	2	O	1 010.64
440	JHB	103.2	97.0	99	1	N	1 042.42
441	JHB	150.0	145.0	226	1	O	663.72
442	JHB	126.0	118.0	108	1	N	1 166.67
443	JHB	125.0	118.0	141	1	O	886.52
444	JHB	115.0	105.0	110	1	O	1 045.45
445	JHB	150.0	124.0	144	1	O	1 041.67
446	JHB	148.0	131.0	130	1	N	1 138.46
447	JHB	150.6	143.0	96	1	N	1 568.75
448	JHB	250.0	215.0	208	1	O	1 201.92
449	JHB	145.0	130.0	139	1	O	1 043.17
450	JHB	138.0	132.0	128	1	O	1 078.13

No	City	Price	Value	Area	Size	Age	PSM
451	JHB	136.0	136.0	109	3	N	1 247.71
452	JHB	166.0	138.0	121	1	O	1 371.90
453	JHB	120.0	115.0	110	1	O	1 090.91
454	JHB	120.0	115.0	130	2	O	923.08
455	JHB	162.0	152.0	134	1	O	1 208.96
456	JHB	88.0	122.0	109	1	O	807.34
457	JHB	105.0	90.0	138	1	O	760.87
458	JHB	116.0	105.0	90	1	O	1 288.89
459	JHB	136.0	123.0	156	1	O	871.79
460	JHB	175.0	172.0	173	2	O	1 011.56
461	JHB	243.0	233.0	198	1	O	1 227.27
462	JHB	161.8	150.0	131	2	N	1 235.11
463	JHB	155.0	143.0	201	1	O	771.14
464	JHB	47.0	83.0	215	3	O	218.60
465	JHB	65.0	87.0	135	2	O	481.48
466	JHB	225.0	199.0	220	1	O	1 022.73
467	JHB	30.0	225.0	217	3	O	138.25
468	JHB	215.0	191.0	191	2	O	1 125.65
469	JHB	280.0	238.0	230	2	O	1 217.39
470	JHB	132.0	123.0	224	1	O	589.29
471	JHB	105.0	95.0	107	1	N	981.31
472	JHB	90.0	78.0	133	1	O	676.69
473	JHB	135.0	130.0	88	1	O	1 534.09
474	JHB	137.0	127.0	217	1	O	631.34
475	JHB	150.0	146.0	146	1	O	1 027.40
476	JHB	230.0	230.0	168	3	O	1 369.05
477	JHB	191.0	185.0	175	1	O	1 091.43
478	JHB	180.0	177.0	218	1	O	825.69
479	JHB	197.0	178.0	146	3	O	1 349.32
480	JHB	70.0	79.0	135	1	O	518.52
481	JHB	265.0	225.0	231	1	O	1 147.19
482	JHB	120.0	115.0	133	2	N	902.26
483	JHB	156.0	131.0	182	2	O	857.14
484	JHB	70.0	68.5	87	1	O	804.60
485	JHB	320.0	268.0	255	2	O	1 254.90
486	JHB	40.0	38.0	100	1	O	400.00
487	JHB	78.0	62.0	83	1	O	939.76
488	JHB	62.0	58.5	89	1	O	696.63
489	JHB	82.0	79.0	109	1	O	752.29
490	JHB	55.0	41.0	96	1	O	572.92
491	JHB	125.0	110.0	163	1	O	766.87
492	JHB	104.0	98.0	132	1	O	787.88
493	JHB	132.0	121.0	151	1	O	874.17
494	JHB	170.0	134.0	218	2	O	779.82
495	JHB	170.0	134.0	218	2	O	779.82
496	JHB	133.0	126.0	167	2	O	796.41
497	JHB	130.0	106.0	153	2	O	849.67
498	JHB	195.0	179.0	181	1	O	1 077.35
499	JHB	65.0	61.0	186	2	O	349.46
500	JHB	27.0	179.0	188	2	O	143.62

No	City	Price	Value	Area	Size	Age	PSM
501	JHB	280.0	241.0	313	2	O	894.57
502	JHB	269.0	165.0	191	2	O	1 408.38
503	JHB	160.0	120.0	111	1	O	1 441.44
504	JHB	302.0	254.0	225	3	O	1 342.22
505	JHB	150.0	144.0	294	1	O	510.20
506	JHB	260.0	233.0	233	1	O	1 115.88
507	JHB	230.0	210.0	175	1	O	1 314.29
508	JHB	335.0	304.0	255	3	O	1 313.73
509	JHB	123.0	230.0	254	3	O	484.25
510	JHB	185.0	162.0	164	2	O	1 128.05
511	JHB	450.0	425.0	272	3	O	1 654.41
512	JHB	115.0	104.0	146	1	O	787.67
513	JHB	260.0	216.0	261	2	O	996.17
514	JHB	162.5	145.0	218	2	O	745.41
515	JHB	135.0	113.0	166	1	O	813.25
516	JHB	130.0	104.0	155	2	O	838.71
517	JHB	87.0	83.0	121	1	O	719.01
518	JHB	405.0	364.0	310	1	N	1 306.45
519	JHB	195.0	182.0	260	2	O	750.00
520	JHB	170.0	150.0	187	3	O	909.09
521	JHB	160.0	154.0	188	1	O	851.06
522	JHB	300.0	285.0	164	2	N	1 829.27
523	JHB	186.0	178.0	182	1	O	1 021.98
524	JHB	230.0	210.0	244	3	O	942.62
525	JHB	95.0	109.5	208	2	O	456.73
526	JHB	131.0	108.0	128	2	O	1 023.44
527	JHB	107.0	98.0	115	1	O	930.43
528	JHB	74.4	75.0	98	1	O	759.18
529	JHB	80.0	69.5	238	3	O	336.13
530	JHB	47.5	45.5	183	1	O	259.56
531	JHB	80.5	74.0	160	2	O	503.13
532	JHB	110.0	101.0	166	2	O	662.65
533	JHB	84.0	79.0	203	1	O	413.79
534	JHB	250.0	218.5	274	2	O	912.41
535	JHB	130.0	119.0	179	1	O	726.26
536	JHB	93.0	88.0	168	1	O	553.57
537	JHB	50.0	47.5	110	1	O	454.55
538	JHB	165.0	142.0	245	2	O	673.47
539	JHB	178.0	134.0	162	2	O	1 098.77
540	JHB	93.0	85.0	195	1	O	476.92
541	JHB	90.8	85.0	203	1	O	447.29
542	JHB	415.0	313.0	327	2	O	1 269.11
543	JHB	150.0	102.0	168	2	O	892.86
544	JHB	107.0	100.0	143	1	O	748.25
545	JHB	118.0	98.0	168	2	O	702.38
546	JHB	100.0	70.5	186	2	O	537.63
547	JHB	70.0	119.0	159	2	O	440.25
548	JHB	100.0	91.0	167	1	O	598.80
549	JHB	140.0	129.0	174	2	O	804.60
550	JHB	430.0	395.0	294	3	O	1 462.59

No	City	Price	Value	Area	Size	Age	PSM
551	JHB	370.0	320.0	325	1	O	1 138.46
552	JHB	145.0	138.0	175	1	O	828.57
553	JHB	140.0	135.0	260	1	O	538.46
554	JHB	240.0	219.0	228	1	O	1 052.63
555	JHB	80.0	72.0	208	1	O	384.62
556	JHB	200.0	186.0	269	1	O	743.49
557	JHB	85.0	75.0	114	1	O	745.61
558	JHB	462.0	462.0	278	3	N	1 661.87
559	JHB	104.0	84.0	109	1	O	954.13
560	JHB	106.0	102.0	249	2	O	425.70
561	JHB	220.0	183.0	168	1	O	1 309.52
562	JHB	140.0	121.0	159	1	O	880.50
563	JHB	75.0	69.0	208	1	O	360.58
564	JHB	89.4	83.5	188	1	O	475.53
565	JHB	126.3	119.3	81	2	N	1 559.26
566	JHB	225.0	221.0	162	2	O	1 388.89
567	JHB	181.7	174.0	198	2	N	917.68
568	JHB	265.0	244.0	223	2	O	1 188.34
569	JHB	405.0	346.0	298	3	O	1 359.06
570	JHB	285.0	249.0	231	2	O	1 233.77
571	JHB	191.0	189.0	154	1	O	1 240.26
572	JHB	390.0	345.0	288	2	O	1 354.17
573	JHB	130.6	127.6	106	2	N	1 232.08
574	JHB	165.0	140.0	143	2	O	1 153.85
575	JHB	165.0	158.0	93	1	N	1 774.19
576	JHB	213.0	208.0	226	3	O	942.48
577	JHB	320.0	281.0	155	3	O	2 064.52
578	JHB	144.4	141.4	81	2	N	1 782.72
579	JHB	233.5	229.5	173	2	N	1 349.71
580	JHB	216.0	211.5	149	2	N	1 449.66
581	JHB	120.0	117.0	120	2	O	1 000.00
582	JHB	335.0	301.0	254	2	O	1 318.90
583	JHB	470.0	460.0	320	2	N	1 468.75
584	JHB	169.5	165.5	110	1	N	1 540.91
585	JHB	197.0	192.5	153	2	N	1 287.58
586	JHB	122.0	150.0	160	2	O	762.50
587	JHB	315.0	285.0	203	3	O	1 551.72
588	JHB	321.4	307.0	192	1	N	1 673.96
589	JHB	164.3	165.0	121	2	N	1 357.85
590	JHB	157.0	145.0	159	2	O	987.42
591	JHB	145.0	137.0	161	2	O	900.62
592	JHB	138.0	130.0	128	1	O	1 078.13
593	JHB	115.0	108.0	118	2	O	974.58
594	JHB	245.0	230.0	203	3	O	1 206.90
595	JHB	165.0	129.0	142	2	N	1 161.97
596	JHB	130.0	127.0	106	2	O	1 226.42
597	JHB	130.0	128.0	227	2	O	572.69
598	JHB	236.0	221.0	193	1	N	1 222.80
599	JHB	188.0	154.0	165	2	O	1 139.39
600	JHB	178.0	175.0	110	2	N	1 618.18

No	City	Price	Value	Area	Size	Age	PSM
601	JHB	216.5	190.0	180	2	O	1 202.78
602	JHB	180.0	204.0	167	1	O	1 077.84
603	JHB	188.0	163.0	188	2	O	1 000.00
604	JHB	175.0	135.0	197	2	O	888.32
605	JHB	138.0	137.0	140	2	N	985.71
606	JHB	230.0	193.0	269	2	O	855.02
607	JHB	265.0	251.0	262	2	N	1 011.45
608	JHB	158.0	148.0	152	2	O	1 039.47
609	JHB	100.0	98.0	136	1	O	735.29
610	JHB	374.0	402.5	211	1	O	1 772.51
611	JHB	180.0	160.0	243	3	O	740.74
612	JHB	100.0	93.0	125	1	O	800.00
613	JHB	125.0	122.0	220	1	O	568.18
614	JHB	95.0	90.0	166	2	O	572.29
615	JHB	110.0	105.0	155	1	O	709.68
616	JHB	210.0	205.0	286	1	O	734.27
617	JHB	53.0	52.0	117	1	O	452.99
618	JHB	95.0	86.0	139	1	O	683.45
619	JHB	162.5	139.0	123	1	O	1 321.14
620	JHB	105.0	98.0	182	2	O	576.92
621	JHB	45.9	58.0	82	1	O	559.76
622	JHB	165.0	153.0	224	1	O	736.61
623	JHB	64.0	87.0	112	1	O	571.43
624	JHB	122.5	110.0	171	1	O	716.37
625	JHB	170.0	188.0	232	2	O	732.76
626	JHB	122.5	109.0	179	1	O	684.36
627	JHB	70.0	68.0	106	1	O	660.38
628	JHB	170.0	150.0	214	1	O	794.39
629	JHB	102.0	92.0	168	1	O	607.14
630	JHB	110.0	105.0	147	1	N	748.30
631	JHB	118.0	108.0	135	1	O	874.07
632	JHB	165.0	155.0	134	1	O	1 231.34
633	JHB	117.0	111.0	95	1	N	1 231.58
634	JHB	300.0	252.0	212	1	O	1 415.09
635	JHB	370.0	325.0	204	2	O	1 813.73
636	JHB	127.0	112.0	131	1	O	969.47
637	JHB	130.0	113.0	115	1	O	1 130.43
638	JHB	195.0	180.0	203	1	O	960.59
639	JHB	60.0	72.0	176	1	O	340.91
640	JHB	285.0	225.0	172	1	O	1 656.98
641	JHB	183.0	171.0	173	1	O	1 057.80
642	JHB	330.0	317.0	206	2	O	1 601.94
643	JHB	296.0	264.0	212	2	O	1 396.23
644	JHB	155.0	140.0	181	1	O	856.35
645	JHB	145.0	135.0	143	1	O	1 013.99
646	JHB	152.0	145.0	114	1	O	1 333.33
647	JHB	350.0	319.0	266	1	O	1 315.79
648	JHB	130.0	111.0	195	2	O	666.67
649	JHB	215.0	195.0	203	1	O	1 059.11
650	JHB	185.0	157.0	254	2	O	728.35

No	City	Price	Value	Area	Size	Age	PSM
651	JHB	285.0	260.0	240	1	O	1 187.50
652	JHB	177.0	142.0	158	2	O	1 120.25
653	JHB	302.0	295.0	265	2	O	1 139.62
654	JHB	79.0	74.0	111	1	O	711.71
655	JHB	285.0	210.0	252	2	O	1 130.95
656	JHB	158.0	135.0	194	1	O	814.43
657	JHB	70.0	68.0	108	1	O	648.15
658	JHB	210.0	200.0	154	1	N	1 363.64
659	JHB	245.0	205.0	211	3	O	1 161.14
660	JHB	315.0	282.0	245	2	O	1 285.71
661	JHB	375.0	335.0	252	2	O	1 488.10
662	JHB	268.0	245.0	296	3	O	905.41
663	JHB	490.0	464.0	336	2	O	1 458.33
664	JHB	310.0	285.0	251	3	O	1 235.06
665	JHB	380.0	334.0	236	2	O	1 610.17
666	JHB	250.0	219.0	192	1	O	1 302.08
667	JHB	126.0	123.0	97	2	N	1 298.97
668	JHB	228.0	200.0	224	3	O	1 017.86
669	JHB	215.0	198.0	272	2	O	790.44
670	JHB	51.0	71.0	131	1	O	389.31
671	JHB	120.0	118.0	196	1	O	612.24
672	JHB	75.0	71.0	114	1	O	657.89
673	JHB	65.0	61.0	129	1	O	503.88
674	JHB	180.0	157.0	195	2	O	923.08
675	JHB	70.0	66.0	184	1	O	380.43
676	JHB	88.0	61.0	160	1	O	550.00
677	JHB	190.0	170.0	232	2	O	818.97
678	JHB	86.0	68.0	95	1	O	905.26
679	JHB	69.8	64.0	134	1	O	520.90
680	JHB	95.0	115.0	139	2	O	683.45
681	JHB	115.0	106.0	176	2	O	653.41
682	JHB	148.0	134.0	193	2	O	766.84
683	JHB	230.0	193.0	323	2	O	712.07
684	JHB	120.0	104.0	200	2	O	600.00
685	JHB	130.0	122.0	231	1	O	562.77
686	JHB	200.0	193.0	307	1	O	651.47
687	JHB	125.0	118.0	117	1	O	1 068.38
688	JHB	245.0	215.0	240	2	O	1 020.83
689	JHB	140.0	132.0	281	1	O	498.22
690	JHB	140.0	133.0	257	2	O	544.75
691	JHB	172.0	160.0	230	2	O	747.83
692	JHB	190.0	179.0	193	2	O	984.46
693	JHB	148.0	132.0	144	1	O	1 027.78
694	JHB	210.0	206.0	316	2	O	664.56
695	JHB	322.1	304.1	235	3	N	1 370.64
696	JHB	460.0	435.0	344	2	N	1 337.21
697	JHB	130.0	196.0	183	2	O	710.38
698	JHB	360.0	333.0	288	2	O	1 250.00
699	JHB	250.0	237.0	273	3	O	915.75
700	JHB	150.0	120.0	115	1	O	1 304.35

No	City	Price	Value	Area	Size	Age	PSM
701	JHB	132.5	125.5	80	2	N	1 656.25
702	JHB	115.2	108.2	89	2	O	1 294.38
703	JHB	132.3	129.3	99	2	N	1 336.36
704	JHB	107.5	104.5	106	2	O	1 014.15
705	JHB	103.0	265.0	261	2	O	394.64
706	JHB	153.0	146.0	183	2	O	836.07
707	JHB	215.0	213.0	141	2	N	1 524.82
708	JHB	188.0	155.0	284	3	O	661.97
709	JHB	135.0	135.0	136	1	N	992.65
710	JHB	192.5	176.0	152	1	N	1 266.45
711	JHB	170.0	144.0	190	2	O	894.74
712	JHB	125.0	122.0	137	2	N	912.41
713	JHB	110.0	100.0	264	2	O	416.67
714	JHB	162.0	139.0	176	2	O	920.45
715	JHB	285.0	280.0	222	2	N	1 283.78
716	JHB	210.1	205.1	188	2	N	1 117.55
717	JHB	105.0	90.0	104	2	O	1 009.62
718	JHB	390.0	360.0	305	2	O	1 278.69
719	JHB	165.0	151.0	150	2	O	1 100.00
720	JHB	125.0	99.0	88	1	O	1 420.45
721	JHB	170.0	143.0	175	1	O	971.43
722	JHB	127.0	114.0	115	1	O	1 104.35
723	JHB	95.0	90.0	130	1	O	730.77
724	JHB	110.0	106.0	182	1	O	604.40
725	JHB	270.0	180.0	172	1	O	1 569.77
726	JHB	80.0	77.0	130	1	O	615.38
727	JHB	93.0	90.0	137	1	O	678.83
728	JHB	80.0	79.0	184	2	O	434.78
729	JHB	76.0	73.0	148	1	O	513.51
730	JHB	132.0	118.0	173	1	O	763.01
731	JHB	90.0	87.0	125	1	N	720.00
732	JHB	200.0	167.0	198	1	O	1 010.10
733	JHB	106.0	101.0	111	1	O	954.95
734	JHB	110.0	86.0	100	1	O	1 100.00
735	JHB	107.0	100.0	80	1	N	1 337.50
736	JHB	114.0	104.0	80	1	N	1 425.00
737	JHB	129.0	122.0	95	1	N	1 357.89
738	JHB	90.0	83.0	86	1	O	1 046.51
739	JHB	66.0	65.0	90	2	O	733.33
740	JHB	85.0	82.0	230	3	O	369.57
741	JHB	148.0	137.0	148	2	O	1 000.00
742	JHB	83.0	62.0	158	1	O	525.32
743	JHB	187.0	182.0	166	1	O	1 126.51
744	JHB	100.0	226.0	214	1	O	467.29
745	JHB	88.0	78.0	146	1	O	602.74
746	JHB	125.0	108.5	173	1	O	722.54
747	JHB	40.0	57.0	95	1	O	421.05
748	JHB	80.0	77.0	91	1	O	879.12
749	JHB	150.0	115.0	181	2	O	828.73
750	JHB	58.0	79.0	127	2	O	456.69

No	City	Price	Value	Area	Size	Age	PSM
751	JHB	157.3	154.3	112	2	N	1 404.46
752	JHB	238.0	233.0	161	2	N	1 478.26
753	JHB	250.0	215.0	208	1	O	1 201.92
754	JHB	500.0	445.0	280	1	O	1 785.71
755	JHB	170.0	153.0	168	1	O	1 011.90
756	JHB	153.0	134.0	156	2	O	980.77
757	JHB	155.0	168.0	182	1	O	851.65
758	JHB	105.0	97.0	204	1	O	514.71
759	JHB	185.0	157.0	254	2	O	728.35
760	JHB	130.0	120.0	140	2	O	928.57
761	JHB	110.0	105.0	160	1	O	687.50
762	JHB	115.0	93.0	122	2	O	942.62
763	JHB	45.5	42.5	165	1	O	275.76
764	JHB	195.0	253.0	290	3	O	672.41
765	JHB	49.0	47.0	94	1	O	521.28
766	JHB	70.0	80.0	176	1	O	397.73
767	JHB	74.0	71.5	89	1	O	831.46
768	JHB	40.0	36.0	290	1	O	137.93
769	JHB	70.0	65.0	108	1	O	648.15
770	JHB	120.0	108.0	169	1	O	710.06
771	JHB	57.0	51.0	121	1	O	471.07
772	JHB	71.0	69.5	90	1	O	788.89
773	JHB	26.5	141.0	252	2	O	105.16
774	JHB	157.5	145.5	182	2	O	865.38
775	JHB	275.0	261.0	254	3	O	1 082.68
776	JHB	240.0	207.0	272	3	O	882.35
777	JHB	240.0	207.0	272	3	O	882.35
778	JHB	95.0	232.0	224	2	O	424.11
779	JHB	62.0	116.0	137	2	O	452.55
780	JHB	80.0	76.0	103	1	O	776.70
781	JHB	45.0	153.0	224	1	O	200.89
782	JHB	325.0	298.0	229	2	N	1 419.21
783	JHB	116.0	99.0	90	1	O	1 288.89
784	JHB	91.0	87.0	161	1	O	565.22
785	JHB	141.0	137.0	178	1	O	792.13
786	JHB	145.0	129.0	131	2	O	1 106.87
787	JHB	105.0	138.0	172	2	O	610.47
788	JHB	236.0	206.0	233	2	O	1 012.88
789	JHB	265.0	245.0	242	2	O	1 095.04
790	JHB	260.0	228.0	184	1	O	1 413.04
791	JHB	225.0	200.0	187	1	O	1 203.21
792	JHB	320.0	280.0	323	3	O	990.71
793	JHB	220.4	220.5	220	1	N	1 001.82
794	JHB	124.0	122.0	100	2	O	1 240.00
795	JHB	185.0	159.0	212	2	O	872.64
796	JHB	230.0	212.0	213	2	O	1 079.81
797	JHB	200.0	193.0	247	2	O	809.72
798	JHB	260.0	233.0	183	1	O	1 420.77
799	JHB	135.0	192.0	193	1	O	699.48
800	JHB	85.0	81.5	160	1	O	531.25

1220 - 1653

No	City	Price	Value	Area	Size	Age	PSM
801	JHB	165.0	152.0	231	2	O	714.29
802	JHB	86.0	79.0	202	1	O	425.74
803	JHB	135.0	120.0	159	2	O	849.06
804	JHB	110.0	93.0	123	1	O	894.31
805	JHB	90.0	86.0	121	1	O	743.80
806	JHB	100.0	97.0	258	1	O	387.60
807	JHB	94.0	112.0	292	2	O	321.92
808	JHB	60.0	81.0	106	1	O	566.04
809	JHB	145.0	132.0	231	2	O	627.71
810	JHB	160.0	152.0	150	2	O	1 066.67
811	JHB	122.0	104.0	85	2	O	1 435.29
812	JHB	108.0	102.0	180	2	O	600.00
813	JHB	124.0	166.0	186	2	O	666.67
814	JHB	90.0	82.0	172	1	O	523.26
815	JHB	60.0	56.0	155	1	O	387.10
816	JHB	100.0	94.0	258	1	O	387.60
817	JHB	57.0	56.0	95	1	O	600.00
818	JHB	88.0	83.0	142	1	O	619.72
819	JHB	96.0	125.0	183	2	O	524.59
820	JHB	82.0	79.5	106	1	O	773.58
821	JHB	183.0	166.0	228	2	N	802.63
822	JHB	260.0	221.0	221	2	O	1 176.47
823	JHB	100.0	96.0	200	1	O	500.00
824	JHB	130.0	119.0	139	2	O	935.25
825	JHB	140.0	239.0	326	2	O	429.45
826	JHB	117.0	111.0	160	1	O	731.25
827	JHB	124.0	106.0	101	2	O	1 227.72
828	JHB	77.0	72.0	161	1	O	478.26
829	JHB	160.0	152.0	213	2	N	751.17
830	JHB	104.0	97.0	135	1	O	770.37
831	JHB	125.0	117.0	198	2	O	631.31
832	JHB	155.0	130.0	194	2	O	798.97
833	JHB	85.0	76.0	159	1	O	534.59
834	JHB	70.0	67.0	128	1	O	546.88
835	JHB	454.0	420.5	339	2	N	1 339.23
836	JHB	275.0	162.0	303	2	O	907.59
837	JHB	165.0	158.0	197	1	O	837.56
838	JHB	149.0	132.0	164	2	O	908.54
839	JHB	82.5	79.0	179	1	O	460.89
840	JHB	220.0	189.0	228	1	O	964.91
841	JHB	165.0	160.0	202	2	O	816.83
842	JHB	130.0	125.0	237	1	O	548.52
843	JHB	120.0	111.0	176	1	O	681.82

No	City	Price	Value	Area	Size	Age	PSM
1	PTA	123.0	106.0	181	2	O	679.56
2	PTA	150.0	144.0	222	3	O	675.68
3	PTA	130.0	122.0	202	2	O	643.56
4	PTA	68.0	79.0	142	1	O	478.87
5	PTA	98.0	106.0	200	1	O	490.00
6	PTA	67.0	58.0	124	1	O	540.32
7	PTA	70.0	67.0	167	1	O	419.16
8	PTA	98.5	94.5	151	2	O	652.32
9	PTA	70.0	111.0	281	2	O	249.11
10	PTA	140.0	132.0	205	2	O	682.93
11	PTA	146.0	108.0	202	2	O	722.77
12	PTA	154.0	153.0	191	2	O	806.28
13	PTA	215.0	199.0	240	2	O	895.83
14	PTA	82.0	106.0	156	2	O	525.64
15	PTA	216.0	184.5	214	2	O	1 009.35
16	PTA	233.0	212.5	170	1	N	1 370.59
17	PTA	144.0	118.0	127	2	O	1 133.86
18	PTA	325.0	249.0	311	3	O	1 045.02
19	PTA	86.0	86.0	84	2	N	1 023.81
20	PTA	87.0	87.0	96	2	O	906.25
21	PTA	180.0	163.0	308	3	O	584.42
22	PTA	131.0	131.0	145	2	N	903.45
23	PTA	105.0	104.0	99	2	O	1 060.61
24	PTA	145.0	140.0	239	2	O	606.69
25	PTA	300.0	283.0	260	3	O	1 153.85
26	PTA	106.0	102.0	132	2	O	803.03
27	PTA	97.0	92.0	93	3	O	1 043.01
28	PTA	167.9	148.8	199	2	N	843.72
29	PTA	98.1	94.1	92	2	O	1 066.30
30	PTA	145.5	145.5	253	2	N	575.10
31	PTA	135.0	120.0	131	2	O	1 030.53
32	PTA	155.0	155.0	187	2	N	828.88
33	PTA	106.0	100.0	102	2	O	1 039.22
34	PTA	120.0	113.5	166	2	O	722.89
35	PTA	127.0	113.0	186	3	O	682.80
36	PTA	115.0	110.5	147	1	O	782.31
37	PTA	209.0	195.0	249	3	O	839.36
38	PTA	165.0	143.0	225	2	O	733.33
39	PTA	101.0	98.0	246	2	O	410.57
40	PTA	94.0	90.0	87	2	O	1 080.46
41	PTA	149.0	137.0	214	2	O	696.26
42	PTA	156.5	155.0	195	2	O	802.56
43	PTA	126.0	119.0	183	2	O	688.52
44	PTA	202.9	171.2	231	2	N	878.35
45	PTA	207.7	198.0	241	2	N	861.83
46	PTA	130.0	125.0	169	3	O	769.23
47	PTA	170.0	147.0	205	2	O	829.27
48	PTA	110.0	110.0	175	2	O	628.57
49	PTA	140.0	135.0	189	2	O	740.74
50	PTA	182.1	181.0	188	2	N	968.62

No	City	Price	Value	Area	Size	Age	PSM
51	PTA	69.0	65.0	110	2	O	627.27
52	PTA	126.0	120.0	164	2	O	768.29
53	PTA	73.8	63.8	94	1	O	785.11
54	PTA	155.0	151.0	225	2	O	688.89
55	PTA	180.5	159.0	258	2	O	699.61
56	PTA	90.0	89.0	142	2	O	633.80
57	PTA	116.0	111.0	94	2	O	1 234.04
58	PTA	238.0	238.0	338	3	N	704.14
59	PTA	108.0	104.5	158	2	O	683.54
60	PTA	135.0	119.0	291	2	O	463.92
61	PTA	105.0	96.0	131	1	O	801.53
62	PTA	110.0	104.0	117	2	O	940.17
63	PTA	115.0	111.0	189	3	O	608.47
64	PTA	124.8	121.3	123	3	N	1 014.63
65	PTA	103.5	96.5	101	2	O	1 024.75
66	PTA	270.0	270.0	286	2	N	944.06
67	PTA	136.3	135.0	112	2	N	1 216.96
68	PTA	132.0	122.0	127	2	O	1 039.37
69	PTA	80.0	80.0	85	2	N	941.18
70	PTA	93.0	81.0	168	2	O	553.57
71	PTA	165.0	161.0	314	2	O	525.48
72	PTA	212.0	199.0	233	2	O	909.87
73	PTA	290.0	262.0	262	3	O	1 106.87
74	PTA	318.0	304.0	216	3	O	1 472.22
75	PTA	159.0	151.0	158	2	N	1 006.33
76	PTA	155.0	129.0	163	2	O	950.92
77	PTA	136.0	125.0	130	2	N	1 046.15
78	PTA	230.0	219.0	285	2	N	807.02
79	PTA	298.8	280.8	342	2	N	873.68
80	PTA	130.0	119.0	209	2	O	622.01
81	PTA	169.0	150.0	191	2	O	884.82
82	PTA	115.0	100.4	81	1	N	1 419.75
83	PTA	220.0	207.0	221	2	O	995.48
84	PTA	166.0	151.0	141	1	N	1 177.30
85	PTA	135.0	121.0	95	1	N	1 421.05
86	PTA	209.0	195.0	249	3	O	839.36
87	PTA	88.5	76.5	163	1	O	542.94
88	PTA	159.7	159.7	141	2	N	1 132.62
89	PTA	139.0	131.0	274	2	O	507.30
90	PTA	92.0	86.0	190	2	O	484.21
91	PTA	125.0	109.0	244	2	O	512.30
92	PTA	68.0	64.0	129	2	O	527.13
93	PTA	70.0	66.0	152	2	O	460.53
94	PTA	51.5	49.0	123	2	O	418.70
95	PTA	79.0	72.0	153	1	O	516.34
96	PTA	81.0	67.0	120	2	O	675.00
97	PTA	74.0	128.0	193	2	O	383.42
98	PTA	75.0	61.0	143	2	O	524.48
99	PTA	144.0	128.0	241	2	O	597.51
100	PTA	210.0	173.0	206	2	O	1 019.42

No	City	Price	Value	Area	Size	Age	PSM
101	PTA	90.0	86.0	137	2	O	656.93
102	PTA	50.0	47.0	265	3	O	188.68
103	PTA	70.0	65.0	111	1	O	630.63
104	PTA	132.0	123.0	222	2	O	594.59
105	PTA	74.0	72.0	91	1	O	813.19
106	PTA	80.0	76.5	168	2	O	476.19
107	PTA	70.2	66.2	178	1	O	394.38
108	PTA	160.0	160.0	247	2	N	647.77
109	PTA	98.0	83.0	166	1	O	590.36
110	PTA	200.0	151.0	261	2	O	766.28
111	PTA	90.0	90.0	231	1	N	389.61
112	PTA	96.5	88.0	89	2	O	1 084.27
113	PTA	105.0	101.0	123	2	O	853.66
114	PTA	105.0	86.0	212	2	O	495.28
115	PTA	116.0	86.0	181	2	O	640.88
116	PTA	187.0	162.0	284	2	O	658.45
117	PTA	141.0	147.0	224	2	O	629.46
118	PTA	60.0	59.5	140	1	O	428.57
119	PTA	90.9	78.9	232	1	O	391.81
120	PTA	159.0	147.0	274	2	O	580.29
121	PTA	80.0	77.0	203	1	O	394.09
122	PTA	90.0	83.0	156	3	O	576.92
123	PTA	72.5	69.5	132	1	O	549.24
124	PTA	135.0	125.0	244	2	O	553.28
125	PTA	60.0	59.0	137	1	O	437.96
126	PTA	84.0	92.4	206	2	O	407.77
127	PTA	238.0	209.0	244	2	O	975.41
128	PTA	209.0	181.0	206	2	O	1 014.56
129	PTA	140.0	132.0	266	2	N	526.32
130	PTA	180.0	168.0	174	2	O	1 034.48
131	PTA	145.5	125.5	163	3	O	892.64
132	PTA	230.0	222.0	295	3	O	779.66
133	PTA	185.0	162.0	258	3	O	717.05
134	PTA	139.0	106.0	255	3	O	545.10
135	PTA	95.0	84.0	144	1	O	659.72
136	PTA	164.0	156.0	228	2	O	719.30
137	PTA	250.0	238.0	296	2	O	844.59
138	PTA	111.0	100.0	217	2	O	511.52
139	PTA	165.6	165.6	219	2	N	756.16
140	PTA	130.0	190.0	239	2	O	543.93
141	PTA	94.5	89.5	221	2	O	427.60
142	PTA	110.0	95.0	94	2	O	1 170.21
143	PTA	172.0	149.0	222	2	O	774.77
144	PTA	205.0	160.0	269	3	O	762.08
145	PTA	269.0	234.0	303	2	N	887.79
146	PTA	106.0	101.5	120	2	O	883.33
147	PTA	112.0	87.0	145	2	O	772.41
148	PTA	80.0	66.0	120	2	O	666.67
149	PTA	94.5	91.5	144	2	O	656.25
150	PTA	220.0	176.5	273	2	N	805.86

No	City	Price	Value	Area	Size	Age	PSM
151	PTA	219.8	183.0	267	2	N	823.22
152	PTA	96.0	96.0	84	2	N	1 142.86
153	PTA	90.0	86.0	158	1	O	569.62
154	PTA	156.0	156.0	161	2	N	968.94
155	PTA	88.0	81.5	112	2	O	785.71
156	PTA	175.0	159.0	273	2	O	641.03
157	PTA	125.0	107.0	197	2	O	634.52
158	PTA	137.0	119.0	187	2	O	732.62
159	PTA	140.0	118.0	256	2	O	546.87
160	PTA	110.0	100.0	161	2	O	683.23
161	PTA	175.0	154.0	223	2	O	784.75
162	PTA	101.0	93.0	89	1	O	1 134.83
163	PTA	140.0	97.0	200	2	O	700.00
164	PTA	206.5	195.5	345	2	N	598.55
165	PTA	127.0	113.0	113	2	O	1 123.89
166	PTA	190.0	173.0	230	2	O	826.09
167	PTA	251.8	229.8	182	1	N	1 383.52
168	PTA	95.0	82.0	194	2	O	489.69
169	PTA	90.0	87.5	120	2	O	750.00
170	PTA	85.0	80.0	219	2	O	388.13
171	PTA	95.0	86.0	145	2	O	655.17
172	PTA	102.0	91.0	174	2	O	586.21
173	PTA	130.0	123.0	185	2	O	702.70
174	PTA	71.0	65.0	186	2	O	381.72
175	PTA	150.0	139.0	248	2	O	604.84
176	PTA	85.0	82.0	155	2	O	548.39
177	PTA	116.0	100.0	183	2	O	633.88
178	PTA	108.0	129.0	268	2	O	402.99
179	PTA	94.0	78.0	231	1	O	406.93
180	PTA	167.5	128.0	204	2	O	821.08
181	PTA	88.0	86.0	112	1	O	785.71
182	PTA	170.0	121.0	219	2	O	776.26
183	PTA	401.2	292.0	318	2	N	1 261.64
184	PTA	88.0	75.0	167	2	O	526.95
185	PTA	115.0	103.0	130	2	O	884.62
186	PTA	126.0	155.0	257	2	O	490.27
187	PTA	130.0	111.5	201	2	O	646.77
188	PTA	90.0	82.0	197	2	O	456.85
189	PTA	70.0	64.5	111	1	O	630.63
190	PTA	145.0	132.0	160	2	O	906.25
191	PTA	390.0	362.5	328	1	N	1 189.02
192	PTA	100.0	138.0	197	3	O	507.61
193	PTA	235.0	213.0	158	1	N	1 487.34
194	PTA	210.0	179.0	240	2	O	875.00
195	PTA	85.0	146.0	243	2	O	349.79
196	PTA	92.5	89.0	98	2	O	943.88
197	PTA	133.5	127.5	158	2	O	844.94
198	PTA	110.0	101.0	96	2	O	1 145.83
199	PTA	108.0	101.5	160	2	O	675.00
200	PTA	187.0	184.0	225	2	O	831.11

No	City	Price	Value	Area	Size	Age	PSM
201	PTA	215.0	210.0	242	2	O	888.43
202	PTA	160.0	142.0	184	2	O	869.57
203	PTA	138.0	131.0	213	2	O	647.89
204	PTA	88.6	86.6	87	2	N	1 018.39
205	PTA	86.0	83.0	152	1	O	565.79
206	PTA	123.0	113.5	147	2	O	836.73
207	PTA	155.0	151.0	202	2	N	767.33
208	PTA	126.0	120.0	222	3	O	567.57
209	PTA	90.0	87.0	106	2	O	849.06
210	PTA	67.3	66.0	122	2	O	551.64
211	PTA	257.0	239.0	262	2	O	980.92
212	PTA	84.9	84.9	80	2	N	1 061.25
213	PTA	130.0	198.0	283	2	O	459.36
214	PTA	200.0	170.0	225	2	O	888.89
215	PTA	102.6	102.6	107	2	N	958.88
216	PTA	120.0	120.0	158	2	O	759.49
217	PTA	116.0	111.0	152	2	O	763.16
218	PTA	98.8	88.8	161	1	O	613.66
219	PTA	97.0	90.0	170	2	O	570.59
220	PTA	173.0	175.0	348	2	O	497.13
221	PTA	138.0	121.5	117	2	O	1 179.49
222	PTA	100.0	98.0	162	2	O	617.28
223	PTA	300.0	279.5	317	2	N	946.37
224	PTA	153.0	147.0	192	2	O	796.87
225	PTA	142.0	138.0	241	2	O	589.21
226	PTA	148.0	147.0	191	2	O	774.87
227	PTA	107.0	99.0	163	2	O	656.44
228	PTA	100.0	94.0	265	2	O	377.36
229	PTA	130.0	117.0	219	2	O	593.61
230	PTA	166.0	147.4	200	2	N	830.00
231	PTA	120.0	113.0	125	2	O	960.00
232	PTA	135.0	99.0	139	2	O	971.22
233	PTA	100.0	98.0	141	2	O	709.22
234	PTA	150.0	130.0	237	2	O	632.91
235	PTA	249.8	250.0	266	2	N	939.10
236	PTA	155.0	150.0	185	2	O	837.84
237	PTA	153.5	142.5	245	2	O	626.53
238	PTA	105.0	99.0	113	2	O	929.20
239	PTA	116.0	112.0	192	2	O	604.17
240	PTA	79.0	82.6	113	2	O	699.12
241	PTA	135.0	110.0	208	2	O	649.04
242	PTA	145.0	135.0	198	3	O	732.32
243	PTA	165.0	158.0	193	2	O	854.92
244	PTA	200.0	182.0	287	2	O	696.86
245	PTA	200.0	182.0	285	2	O	701.75
246	PTA	185.0	152.0	199	2	O	929.65
247	PTA	170.0	154.5	253	2	O	671.94
248	PTA	100.0	100.0	117	2	N	854.70
249	PTA	120.0	108.0	84	2	O	1 428.57
250	PTA	190.0	182.0	298	3	O	637.58

No	City	Price	Value	Area	Size	Age	PSM
251	PTA	322.0	310.0	331	2	N	972.81
252	PTA	82.0	72.0	93	2	O	881.72
253	PTA	218.0	176.0	149	2	O	1 463.09
254	PTA	97.0	85.0	90	1	N	1 077.78
255	PTA	184.0	156.0	227	2	O	810.57
256	PTA	214.0	190.0	315	2	O	679.37
257	PTA	250.0	240.0	312	3	O	801.28
258	PTA	210.0	191.0	202	2	O	1 039.60
259	PTA	90.0	157.0	232	2	O	387.93
260	PTA	165.0	145.8	210	2	O	785.71
261	PTA	100.0	95.0	94	2	N	1 063.83
262	PTA	190.0	174.0	173	3	O	1 098.27
263	PTA	96.0	80.0	208	2	O	461.54
264	PTA	187.0	162.0	284	2	O	658.45
265	PTA	130.0	138.0	171	2	O	760.23
266	PTA	200.0	200.0	206	2	N	970.87
267	PTA	112.0	107.5	91	1	O	1 230.77
268	PTA	90.0	81.0	91	1	N	989.01
269	PTA	150.0	143.0	180	1	O	833.33
270	PTA	85.0	82.0	122	1	O	696.72
271	PTA	96.0	94.0	81	2	O	1 185.19
272	PTA	64.2	57.2	89	1	O	721.35
273	PTA	80.0	80.0	177	2	O	451.98
274	PTA	228.0	195.0	229	2	O	995.63
275	PTA	195.0	184.0	319	2	O	611.29
276	PTA	120.0	108.0	181	2	O	662.98
277	PTA	80.0	75.0	116	2	O	689.66
278	PTA	120.0	103.0	124	2	O	967.74
279	PTA	100.0	85.0	166	2	O	602.41
280	PTA	140.0	103.0	281	2	O	498.22
281	PTA	182.0	161.0	255	2	O	713.73
282	PTA	86.0	154.0	221	3	O	389.14
283	PTA	165.0	146.0	250	2	O	660.00
284	PTA	79.0	68.0	102	2	O	774.51
285	PTA	90.0	83.5	186	2	O	483.87
286	PTA	120.0	121.0	272	1	O	441.18
287	PTA	85.0	81.0	175	2	O	485.71
288	PTA	173.3	150.3	135	1	N	1 283.70
289	PTA	160.0	122.0	186	2	O	860.22
290	PTA	78.0	69.0	179	1	O	435.75
291	PTA	194.0	210.0	282	3	O	687.94
292	PTA	225.0	203.0	230	2	O	978.26
293	PTA	120.0	117.0	238	2	O	504.20
294	PTA	80.0	75.0	96	2	O	833.33
295	PTA	66.0	64.0	125	2	O	528.00
296	PTA	175.0	157.0	206	2	O	849.51
297	PTA	155.0	146.0	171	2	O	906.43
298	PTA	183.0	159.5	217	2	O	843.32
299	PTA	108.0	99.0	159	2	O	679.25
300	PTA	185.0	179.0	232	2	O	797.41

No	City	Price	Value	Area	Size	Age	PSM
301	PTA	165.0	153.0	123	2	O	1 341.46
302	PTA	120.0	120.0	139	2	N	863.31
303	PTA	87.0	87.0	84	2	N	1 035.71
304	PTA	150.0	144.0	176	2	O	852.27
305	PTA	162.5	146.0	140	1	O	1 160.71
306	PTA	170.0	148.0	209	2	O	813.40
307	PTA	147.0	126.0	156	2	O	942.31
308	PTA	130.0	125.0	227	2	O	572.69
309	PTA	265.0	214.0	277	3	O	956.68
310	PTA	129.0	122.0	272	2	O	474.26
311	PTA	100.0	96.0	135	2	O	740.74
312	PTA	97.0	88.0	91	2	O	1 065.93
313	PTA	153.0	155.0	272	3	O	562.50
314	PTA	115.0	109.0	109	2	O	1 055.05
315	PTA	115.0	100.0	125	2	O	920.00
316	PTA	87.0	82.0	143	1	O	608.39
317	PTA	193.6	193.6	251	2	N	771.31
318	PTA	175.0	160.0	277	2	O	631.77
319	PTA	125.0	115.0	134	2	O	932.84
320	PTA	108.0	99.0	148	2	O	729.73
321	PTA	122.0	117.0	105	2	O	1 161.90
322	PTA	122.0	116.0	105	1	O	1 161.90
323	PTA	298.0	270.0	350	2	O	851.43
324	PTA	115.0	105.0	221	2	O	520.36
325	PTA	85.0	82.0	183	2	O	464.48
326	PTA	203.0	199.0	239	3	N	849.37
327	PTA	97.8	97.8	98	2	N	997.96
328	PTA	65.0	58.0	97	2	O	670.10
329	PTA	149.0	140.0	224	2	O	665.18
330	PTA	150.0	129.0	190	2	O	789.47
331	PTA	106.0	98.5	85	2	O	1 247.06
332	PTA	105.0	96.0	98	2	O	1 071.43
333	PTA	109.0	97.0	120	2	O	908.33
334	PTA	345.0	316.0	336	2	O	1 026.79
335	PTA	99.0	96.0	126	2	O	785.71
336	PTA	168.0	138.0	208	2	O	807.69
337	PTA	135.0	121.0	211	2	N	639.81
338	PTA	205.0	180.0	268	2	O	764.93
339	PTA	80.0	72.0	180	2	O	444.44
340	PTA	161.0	156.0	192	2	O	838.54
341	PTA	180.0	145.0	182	2	O	989.01
342	PTA	138.0	132.0	153	2	O	901.96
343	PTA	117.0	115.0	103	2	N	1 135.92
344	PTA	162.5	146.0	140	1	O	1 160.71
345	PTA	128.0	119.0	173	2	O	739.88
346	PTA	174.0	153.0	202	3	O	861.39
347	PTA	111.2	111.2	117	2	N	950.43
348	PTA	265.0	249.0	310	2	N	854.84
349	PTA	132.0	126.5	105	2	N	1 257.14
350	PTA	107.0	104.0	185	2	O	578.38

No	City	Price	Value	Area	Size	Age	PSM
351	PTA	70.0	67.0	93	2	O	752.69
352	PTA	145.0	137.0	152	2	O	953.95
353	PTA	77.6	109.5	170	1	O	456.47
354	PTA	200.6	192.6	203	1	O	988.18
355	PTA	160.0	150.0	200	2	O	800.00
356	PTA	80.0	76.0	171	2	O	467.84
357	PTA	150.0	83.0	146	1	O	1 027.40
358	PTA	200.0	200.0	304	2	O	657.89
359	PTA	115.0	106.0	122	2	O	942.62
360	PTA	220.0	191.0	294	2	O	748.30
361	PTA	257.9	243.9	284	2	N	908.10
362	PTA	116.0	103.0	224	2	O	517.86
363	PTA	96.5	85.5	213	1	O	453.05
364	PTA	195.0	195.0	189	2	N	1 031.75
365	PTA	166.0	149.0	144	2	O	1 152.78
366	PTA	94.0	88.0	200	2	O	470.00
367	PTA	90.0	90.0	87	2	N	1 034.48
368	PTA	135.0	117.0	253	3	O	533.60
369	PTA	101.5	91.5	160	2	O	634.37
370	PTA	104.5	99.0	139	2	O	751.80
371	PTA	226.6	206.6	224	2	N	1 011.61
372	PTA	150.0	141.0	199	2	O	753.77
373	PTA	147.0	133.0	179	2	O	821.23
374	PTA	306.0	279.0	333	3	O	918.92
375	PTA	149.0	141.0	208	2	O	716.35
376	PTA	185.0	166.0	207	2	O	893.72
377	PTA	125.0	107.0	166	2	N	753.01
378	PTA	108.0	128.0	182	2	O	593.41
379	PTA	169.0	152.0	232	3	O	728.45
380	PTA	120.0	110.0	173	2	O	693.64
381	PTA	160.0	137.0	156	2	O	1 025.64
382	PTA	126.0	118.0	194	2	O	649.48
383	PTA	90.0	131.0	222	2	O	405.41
384	PTA	113.3	105.3	87	2	O	1 302.30
385	PTA	162.5	154.5	198	2	O	820.71
386	PTA	190.0	171.0	221	2	O	859.73
387	PTA	105.0	99.0	95	2	O	1 105.26
388	PTA	145.0	130.0	264	2	O	549.24
389	PTA	131.0	110.0	174	2	O	752.87
390	PTA	210.0	200.0	328	3	O	640.24
391	PTA	126.0	120.0	123	2	O	1 024.39
392	PTA	102.6	97.0	85	1	N	1 207.06
393	PTA	176.0	168.0	230	2	O	765.22
394	PTA	170.0	160.0	197	3	O	862.94
395	PTA	135.0	124.0	167	2	O	808.38
396	PTA	125.0	115.0	178	2	O	702.25
397	PTA	121.0	120.0	136	2	N	889.71
398	PTA	146.0	126.0	156	2	O	935.90
399	PTA	218.0	193.0	222	2	O	981.98
400	PTA	118.0	92.0	149	2	O	791.95

No	City	Price	Value	Area	Size	Age	PSM
401	PTA	108.0	102.0	92	2	O	1 173.91
402	PTA	92.6	85.6	183	1	O	506.01
403	PTA	86.0	83.5	110	2	O	781.82
404	PTA	132.0	124.0	188	2	O	702.13
405	PTA	128.0	110.0	165	2	O	775.76
406	PTA	109.5	109.5	122	2	N	897.54
407	PTA	140.0	133.0	183	2	O	765.03
408	PTA	207.0	197.0	251	2	O	824.70
409	PTA	213.0	185.0	288	2	O	739.58
410	PTA	188.0	174.0	284	2	O	661.97
411	PTA	96.0	92.0	166	2	O	578.31
412	PTA	109.0	101.0	85	2	O	1 282.35
413	PTA	150.0	134.0	217	2	O	691.24
414	PTA	135.0	128.0	161	2	O	838.51
415	PTA	135.0	121.0	145	2	O	931.03
416	PTA	265.0	226.0	225	3	O	1 177.78
417	PTA	230.0	210.0	302	3	O	761.59
418	PTA	190.0	162.0	300	2	O	633.33
419	PTA	230.0	175.0	195	2	O	1 179.49
420	PTA	137.0	126.0	128	2	O	1 070.31
421	PTA	172.0	153.0	156	2	N	1 102.56
422	PTA	190.0	170.0	203	2	O	935.96
423	PTA	132.0	153.0	192	2	O	687.50
424	PTA	224.8	195.8	279	2	N	805.73
425	PTA	139.0	126.0	134	2	N	1 037.31
426	PTA	233.0	228.0	232	2	N	1 004.31
427	PTA	204.0	203.0	242	2	N	842.98
428	PTA	220.0	196.0	267	2	O	823.97
429	PTA	95.0	89.0	147	2	O	646.26
430	PTA	192.0	173.0	231	2	O	831.17
431	PTA	139.4	127.9	101	2	N	1 380.20
432	PTA	180.0	171.0	204	2	N	882.35
433	PTA	112.0	97.0	211	2	O	530.81
434	PTA	180.7	175.7	206	2	N	877.18
1	CT	92.2	92.2	94	1	N	980.85
2	CT	150.0	142.0	245	2	O	612.24
3	CT	124.3	119.3	108	2	N	1 150.93
4	CT	105.5	104.5	80	2	N	1 318.75
5	CT	163.9	163.9	216	2	N	758.80
6	CT	117.0	112.0	200	1	O	585.00
7	CT	130.0	125.0	126	1	O	1 031.75
8	CT	108.0	176.0	207	2	O	521.74
9	CT	270.0	255.0	260	2	O	1 038.46
10	CT	186.0	165.0	294	1	O	632.65

No	City	Price	Value	Area	Size	Age	PSM
11	CT	94.5	94.5	80	1	N	1 181.25
12	CT	96.9	96.9	80	1	N	1 211.25
13	CT	52.0	52.0	268	1	O	194.03
14	CT	134.0	127.0	153	2	O	875.82
15	CT	87.0	87.0	101	1	N	861.39
16	CT	83.2	81.2	83	1	O	1 002.41
17	CT	83.4	81.4	123	1	O	678.05
18	CT	98.0	94.0	173	1	O	566.47
19	CT	94.5	91.5	98	1	O	964.29
20	CT	74.0	71.0	115	1	O	643.48
21	CT	215.0	205.0	205	2	O	1 048.78
22	CT	77.0	76.0	102	1	O	754.90
23	CT	91.5	91.5	84	1	N	1 089.29
24	CT	178.0	173.0	142	2	O	1 253.52
25	CT	250.0	230.0	207	2	O	1 207.73
26	CT	107.0	107.0	95	2	N	1 126.32
27	CT	35.0	35.0	106	1	O	330.19
28	CT	210.0	195.0	217	2	O	967.74
29	CT	103.0	83.0	151	1	O	682.12
30	CT	136.0	140.0	180	1	O	755.56
31	CT	165.0	157.0	252	2	O	654.76
32	CT	106.0	96.0	115	1	O	921.74
33	CT	86.0	82.0	80	1	N	1 075.00
34	CT	122.0	110.0	113	1	O	1 079.65
35	CT	87.0	82.0	128	1	O	679.69
36	CT	170.0	162.0	176	2	O	965.91
37	CT	76.8	72.8	138	1	O	556.52
38	CT	135.0	130.0	108	1	O	1 250.00
39	CT	90.0	145.0	208	1	O	432.69
40	CT	76.6	73.6	90	1	O	851.11
41	CT	95.0	90.0	92	1	O	1 032.61
42	CT	90.0	88.0	154	1	O	584.42
43	CT	154.0	154.0	115	3	N	1 339.13
44	CT	96.2	96.2	85	1	N	1 131.76
45	CT	168.7	161.7	163	1	N	1 034.97
46	CT	142.5	138.5	143	2	N	996.50
47	CT	68.0	210.0	222	3	O	306.31
48	CT	114.5	112.5	103	2	N	1 111.65
49	CT	96.0	93.0	105	2	O	914.29
50	CT	80.0	77.0	146	1	O	547.95
51	CT	40.0	165.0	160	1	O	250.00
52	CT	116.8	100.0	120	1	N	973.33
53	CT	56.0	55.0	211	1	O	265.40
54	CT	89.0	155.8	169	1	O	526.63
55	CT	88.0	84.0	93	1	O	946.24
56	CT	300.0	295.0	254	1	O	1 181.10
57	CT	76.2	71.2	129	1	O	590.70
58	CT	110.0	104.0	104	1	O	1 057.69
59	CT	170.0	168.0	145	2	O	1 172.41
60	CT	68.0	65.0	130	1	O	523.08

No	City	Price	Value	Area	Size	Age	PSM
61	CT	95.0	90.0	109	1	O	871.56
62	CT	65.0	59.0	110	1	O	590.91
63	CT	92.5	90.3	112	1	N	825.89
64	CT	91.0	86.0	180	1	O	505.56
65	CT	200.0	380.0	230	2	O	869.57
66	CT	135.0	130.0	136	1	O	992.65
67	CT	85.0	77.0	117	1	O	726.50
68	CT	120.0	270.0	256	2	O	468.75
69	CT	142.0	142.0	186	2	O	763.44
70	CT	102.0	94.0	110	1	O	927.27
71	CT	115.3	108.3	113	1	O	1 020.35
72	CT	155.0	140.0	171	1	O	906.43
73	CT	59.5	50.0	144	1	O	413.19
74	CT	82.0	78.0	162	1	O	506.17
75	CT	67.0	105.0	87	1	O	770.11
76	CT	130.0	110.0	210	1	O	619.05
77	CT	56.0	53.0	93	1	O	602.15
78	CT	169.0	141.0	165	1	O	1 024.24
79	CT	115.0	110.0	162	1	O	709.88
80	CT	380.0	355.0	187	2	O	2 032.09
81	CT	83.9	82.0	102	1	O	822.55
82	CT	150.0	141.0	162	1	O	925.93
83	CT	98.0	120.0	137	1	O	715.33
84	CT	70.0	66.0	90	1	O	777.78
85	CT	58.0	58.5	81	1	O	716.05
86	CT	73.0	71.0	90	1	O	811.11
87	CT	160.0	130.0	148	1	O	1 081.08
88	CT	85.9	83.0	102	1	O	842.16
89	CT	65.0	63.0	216	1	O	300.93
90	CT	70.0	63.0	83	1	O	843.37
91	CT	100.0	153.0	236	1	O	423.73
92	CT	65.0	63.0	101	1	O	643.56
93	CT	85.0	76.7	151	1	O	562.91
94	CT	74.0	70.0	89	1	O	831.46
95	CT	230.0	222.0	227	2	O	1 013.22
96	CT	78.0	78.0	90	1	O	866.67
97	CT	65.0	65.0	96	1	O	677.08
98	CT	218.0	218.0	131	2	N	1 664.12
99	CT	58.2	57.0	89	1	O	653.93
100	CT	202.5	185.0	124	1	O	1 633.06
101	CT	81.5	77.5	140	1	O	582.14
102	CT	87.0	86.0	133	1	O	654.14
103	CT	90.0	90.0	143	1	O	629.37
104	CT	93.5	93.5	139	1	O	672.66
105	CT	90.0	139.1	175	2	O	514.29
106	CT	159.6	159.6	124	2	N	1 287.10
107	CT	275.0	250.0	192	2	O	1 432.29
108	CT	315.0	310.0	344	2	O	915.70
109	CT	90.0	89.0	94	1	O	957.45
110	CT	35.0	35.0	233	2	O	150.21

No	City	Price	Value	Area	Size	Age	PSM
111	CT	150.0	141.0	219	2	O	684.93
112	CT	237.0	217.0	151	1	O	1 569.54
113	CT	125.0	122.0	179	2	O	698.32
114	CT	89.9	81.0	102	1	O	881.37
115	CT	320.0	260.0	214	2	N	1 495.33
116	CT	218.3	218.3	254	2	N	859.45
117	CT	111.0	135.0	160	1	O	693.75
118	CT	60.0	60.0	102	1	O	588.24
119	CT	138.0	121.0	115	1	O	1 200.00
120	CT	80.0	70.0	87	1	O	919.54
121	CT	97.0	94.0	129	2	O	751.94
122	CT	143.7	116.7	141	1	O	1 019.15
123	CT	95.0	88.5	122	1	O	778.69
124	CT	145.0	134.0	150	1	O	966.67
125	CT	155.0	135.0	134	1	O	1 156.72
126	CT	78.0	78.0	100	1	O	780.00
127	CT	110.0	94.6	90	1	O	1 222.22
128	CT	166.0	160.0	188	2	O	882.98
129	CT	115.0	105.0	158	1	O	727.85
130	CT	130.0	125.0	194	1	O	670.10
131	CT	135.0	130.0	122	1	O	1 106.56
132	CT	117.0	103.0	169	1	O	692.31
133	CT	105.0	96.0	95	1	O	1 105.26
134	CT	105.0	101.0	256	1	O	410.16
135	CT	90.0	87.3	130	1	O	692.31
136	CT	90.0	85.0	83	1	O	1 084.34
137	CT	77.0	76.0	80	1	O	962.50
138	CT	86.0	82.0	84	1	O	1 023.81
139	CT	185.0	180.0	237	2	O	780.59
140	CT	164.1	161.5	162	1	N	1 012.96
141	CT	265.0	235.0	125	1	O	2 120.00
142	CT	110.0	110.0	125	2	N	880.00
143	CT	99.0	87.0	148	1	O	668.92
144	CT	28.0	32.0	85	1	O	329.41
145	CT	200.0	200.0	202	1	N	990.10
146	CT	100.0	96.0	104	1	O	961.54
147	CT	155.0	144.0	143	1	O	1 083.92
148	CT	110.0	101.0	263	1	O	418.25
149	CT	70.0	73.0	99	1	O	707.07
150	CT	57.5	91.8	140	1	O	410.71
151	CT	95.5	90.5	124	1	O	770.16
152	CT	70.0	65.0	96	1	O	729.17
153	CT	137.4	137.4	134	2	N	1 025.37
154	CT	260.0	232.0	216	2	O	1 203.70
155	CT	163.0	158.0	137	1	O	1 189.78
156	CT	100.0	94.0	172	1	O	581.40
157	CT	95.0	91.0	171	1	O	555.56
158	CT	80.0	75.0	126	1	O	634.92
159	CT	127.0	122.7	164	1	O	774.39
160	CT	210.0	175.0	167	1	O	1 257.49

No	City	Price	Value	Area	Size	Age	PSM
161	CT	159.0	124.0	141	1	O	1 127.66
162	CT	56.0	97.0	166	1	O	337.35
163	CT	415.0	360.0	118	1	O	3 516.95
164	CT	145.0	117.0	93	1	O	1 559.14
165	CT	78.0	78.0	201	1	O	388.06
166	CT	122.0	105.5	107	1	O	1 140.19
167	CT	125.0	122.0	126	2	O	992.06
168	CT	108.0	104.5	157	1	O	687.90
169	CT	66.5	62.0	100	1	O	665.00
170	CT	125.0	117.5	195	1	O	641.03
171	CT	145.0	173.5	192	1	O	755.21
172	CT	255.0	220.0	170	1	O	1 500.00
173	CT	275.0	335.0	202	2	O	1 361.39
174	CT	325.0	321.0	312	1	N	1 041.67
175	CT	129.0	122.0	92	1	O	1 402.17
176	CT	135.0	130.0	105	1	O	1 285.71
177	CT	120.0	158.0	156	2	O	769.23
178	CT	124.0	114.0	137	1	O	905.11
179	CT	115.0	108.0	176	1	O	653.41
180	CT	107.0	86.0	175	1	O	611.43
181	CT	142.3	141.0	106	2	N	1 342.45
182	CT	149.7	150.0	130	2	N	1 151.54
183	CT	68.0	235.0	329	2	O	206.69
184	CT	69.9	67.4	80	1	O	873.75
185	CT	108.0	103.0	120	2	N	900.00
186	CT	160.0	152.0	138	2	O	1 159.42
187	CT	270.0	261.0	240	2	O	1 125.00
188	CT	146.0	140.0	97	1	O	1 505.15
189	CT	122.0	105.0	99	1	O	1 232.32
190	CT	185.0	180.0	138	1	O	1 340.58
191	CT	177.2	177.2	171	2	N	1 036.26
192	CT	150.0	144.0	137	1	O	1 094.89
193	CT	147.0	142.0	131	1	O	1 122.14
194	CT	105.0	95.0	104	1	O	1 009.62
195	CT	160.0	150.0	213	2	O	751.17
196	CT	200.0	193.0	149	1	O	1 342.28
197	CT	65.0	63.0	103	1	O	631.07
198	CT	206.0	265.0	259	2	O	795.37
199	CT	85.0	81.0	193	2	O	440.41
200	CT	200.0	200.0	173	2	N	1 156.07
201	CT	108.0	95.0	161	1	O	670.81
202	CT	53.0	71.0	129	1	O	410.85
203	CT	120.0	118.0	101	1	O	1 188.12
204	CT	194.0	185.0	198	2	O	979.80
205	CT	91.7	91.7	84	1	N	1 091.67
206	CT	63.5	63.5	86	1	O	738.37
207	CT	141.1	141.1	120	2	N	1 175.83
208	CT	115.0	107.0	166	1	O	692.77
209	CT	74.0	70.0	82	1	O	902.44
210	CT	75.0	73.0	92	1	O	815.22

No	City	Price	Value	Area	Size	Age	PSM
211	CT	42.0	225.0	339	2	O	123.89
212	CT	241.2	241.2	237	2	N	1 017.72
213	CT	78.0	118.0	162	1	O	481.48
214	CT	91.5	91.5	83	1	N	1 102.41
215	CT	116.0	96.0	123	1	O	943.09
216	CT	90.0	87.0	92	2	O	978.26
217	CT	152.0	147.0	166	2	O	915.66
218	CT	176.0	176.0	168	2	N	1 047.62
219	CT	101.5	100.0	82	1	N	1 237.80
220	CT	150.0	142.0	220	1	O	681.82
221	CT	50.0	47.0	106	1	O	471.70
222	CT	97.5	83.0	136	1	O	716.91
223	CT	139.3	134.3	153	1	N	910.46
224	CT	155.0	130.0	197	2	O	786.80
225	CT	69.1	65.1	169	1	O	408.88
226	CT	98.0	84.0	141	1	O	695.04
227	CT	150.0	135.0	169	2	O	887.57
228	CT	121.0	121.0	109	2	N	1 110.09
229	CT	120.5	114.5	113	1	N	1 066.37
230	CT	108.0	224.0	230	2	O	469.57
231	CT	119.0	101.0	144	1	O	826.39
232	CT	157.0	140.0	202	1	O	777.23
233	CT	220.0	195.0	267	2	O	823.97
234	CT	110.0	110.0	116	1	O	948.28
235	CT	103.8	104.0	108	1	N	961.11
236	CT	70.0	68.0	103	1	O	679.61
237	CT	112.0	104.5	133	2	O	842.11
238	CT	320.0	280.0	286	2	O	1 118.88
239	CT	77.0	68.0	86	1	O	895.35
240	CT	126.0	118.0	165	1	O	763.64
241	CT	265.0	230.0	250	2	O	1 060.00
242	CT	203.0	193.0	212	2	O	957.55
243	CT	165.0	165.0	160	1	N	1 031.25
244	CT	122.0	122.0	114	2	N	1 070.18
245	CT	130.5	130.5	114	2	N	1 144.74
246	CT	110.0	105.0	129	1	O	852.71
247	CT	57.0	57.0	83	1	O	686.75
248	CT	131.0	131.0	126	2	N	1 039.68
249	CT	106.0	97.0	164	1	O	646.34
250	CT	91.5	91.5	83	1	N	1 102.41
251	CT	158.0	147.0	174	2	O	908.05
252	CT	112.5	112.5	97	2	N	1 159.79
253	CT	160.0	158.0	220	2	N	727.27
254	CT	237.0	173.0	203	2	O	1 167.49
255	CT	257.8	258.0	260	2	N	991.54
256	CT	125.0	118.0	185	1	O	675.68
257	CT	76.5	74.5	84	1	O	910.71
258	CT	70.0	68.0	94	1	O	744.68
259	CT	140.0	130.0	145	2	O	965.52
260	CT	87.0	83.0	105	1	O	828.57

No	City	Price	Value	Area	Size	Age	PSM
261	CT	250.0	212.0	180	2	O	1 388.89
262	CT	56.6	55.0	190	1	O	297.89
263	CT	160.0	145.0	159	1	O	1 006.29
264	CT	171.0	140.0	98	1	O	1 744.90
265	CT	191.0	163.0	291	2	O	656.36
266	CT	165.0	160.0	160	2	O	1 031.25
267	CT	61.5	150.0	193	2	O	318.65
268	CT	120.0	106.6	103	2	N	1 165.05
269	CT	129.4	129.4	149	2	N	868.46
270	CT	170.0	162.0	141	1	N	1 205.67
271	CT	78.0	240.0	160	3	O	487.50
272	CT	88.0	83.0	90	1	O	977.78
273	CT	135.0	129.0	161	1	O	838.51
274	CT	64.6	60.0	144	1	O	448.61
275	CT	190.0	185.0	150	1	O	1 266.67
276	CT	160.0	148.0	255	2	O	627.45
277	CT	110.0	100.0	161	1	O	683.23
278	CT	205.0	163.0	131	1	O	1 564.89
279	CT	120.0	229.0	241	2	O	497.93
280	CT	145.0	129.0	191	2	O	759.16
281	CT	80.0	79.0	89	1	O	898.88
282	CT	285.0	265.0	143	1	O	1 993.01
283	CT	105.9	100.9	98	1	O	1 080.61
284	CT	62.6	55.8	91	1	O	687.91
285	CT	60.0	58.0	90	1	O	666.67
286	CT	76.0	72.0	148	2	O	513.51
287	CT	59.0	53.0	88	1	O	670.45
288	CT	51.3	51.0	91	1	O	563.74
289	CT	79.0	74.0	127	1	O	622.05
290	CT	72.0	100.0	158	1	O	455.70
291	CT	75.0	73.0	124	1	O	604.84
292	CT	94.5	80.0	109	1	N	866.97
293	CT	77.5	75.0	118	1	O	656.78
294	CT	75.0	70.0	99	1	O	757.58
295	CT	70.0	64.5	80	1	O	875.00
296	CT	70.0	64.5	80	1	O	875.00
297	CT	120.0	95.0	175	1	O	685.71
298	CT	57.5	55.5	89	1	O	646.07
299	CT	117.0	111.0	180	1	O	650.00
300	CT	98.5	120.0	100	1	O	985.00
301	CT	68.0	60.6	82	1	O	829.27
302	CT	212.0	207.0	183	2	O	1 158.47
303	CT	89.0	86.0	122	1	O	729.51
304	CT	94.0	84.0	111	1	O	846.85
305	CT	275.6	267.3	288	2	N	956.94
306	CT	72.0	69.0	85	1	O	847.06
307	CT	76.0	73.0	80	1	O	950.00
308	CT	130.0	75.0	133	1	N	977.44
309	CT	105.5	98.0	136	1	O	775.74
310	CT	65.0	62.0	98	1	O	663.27

No	City	Price	Value	Area	Size	Age	PSM
311	CT	105.0	102.0	150	1	O	700.00
312	CT	55.0	51.0	83	1	O	662.65
313	CT	135.0	125.0	168	1	O	803.57
314	CT	51.0	49.0	80	1	O	637.50
315	CT	76.9	74.0	87	1	O	883.91
316	CT	65.4	63.0	90	1	O	726.67
317	CT	115.0	305.0	302	1	O	380.79
318	CT	58.9	55.9	83	1	O	709.64
319	CT	60.0	50.0	146	3	O	410.96
320	CT	165.0	139.0	337	2	O	489.61
321	CT	115.0	189.0	240	2	O	479.17
322	CT	74.5	69.7	155	2	O	480.65
323	CT	88.5	86.5	116	2	O	762.93
324	CT	88.0	85.6	149	2	O	590.60
325	CT	65.0	65.0	210	2	O	309.52
326	CT	320.0	185.0	301	3	O	1 063.12
327	CT	160.0	127.0	196	1	O	816.33
328	CT	92.0	86.0	154	2	O	597.40
329	CT	63.0	130.3	147	1	O	428.57
330	CT	103.0	95.0	98	1	O	1 051.02
331	CT	145.0	148.2	221	2	O	656.11
332	CT	120.0	107.0	169	1	O	710.06
333	CT	30.0	47.9	221	3	O	135.75
334	CT	31.6	37.0	153	2	O	206.54
335	CT	125.0	123.5	187	2	O	668.45
336	CT	66.0	67.0	113	1	O	584.07
337	CT	110.0	100.0	289	2	O	380.62
338	CT	106.0	103.0	152	1	O	697.37
339	CT	71.0	71.0	116	2	O	612.07
340	CT	62.5	62.5	83	1	O	753.01
341	CT	200.0	197.0	229	2	O	873.36
342	CT	118.0	118.0	102	1	N	1 156.86
343	CT	80.0	80.0	126	1	N	634.92
344	CT	85.0	82.5	107	1	O	794.39
345	CT	202.4	202.4	145	2	N	1 395.86
346	CT	27.5	50.0	88	1	O	312.50
347	CT	185.0	179.0	162	1	O	1 141.98
348	CT	170.0	140.0	152	1	O	1 118.42
349	CT	155.0	125.0	139	1	O	1 115.11
350	CT	168.0	155.0	133	1	O	1 263.16
351	CT	99.0	95.0	156	1	O	634.62
352	CT	68.1	68.1	99	1	N	687.88
353	CT	63.7	57.0	100	1	O	637.00
354	CT	273.0	258.0	247	2	O	1 105.26
355	CT	95.0	86.0	166	1	O	572.29
356	CT	54.0	112.3	201	1	O	268.66
357	CT	67.5	104.0	155	1	O	435.48
358	CT	265.0	250.0	190	1	O	1 394.74
359	CT	196.0	186.0	246	2	O	796.75
360	CT	110.0	102.0	170	1	O	647.06

No	City	Price	Value	Area	Size	Age	PSM
361	CT	90.0	88.0	127	1	O	708.66
362	CT	85.0	80.0	143	1	O	594.41
363	CT	85.0	80.0	102	1	O	833.33
364	CT	117.0	103.0	169	1	O	692.31
365	CT	120.0	110.0	138	1	O	869.57
366	CT	122.0	117.0	221	2	O	552.04
367	CT	117.5	114.0	157	1	O	748.41
368	CT	170.0	142.0	181	1	O	939.23
369	CT	95.0	90.0	124	1	O	766.13
370	CT	84.0	80.5	127	1	O	661.42
371	CT	75.0	180.0	277	2	O	270.76
372	CT	76.0	82.6	139	1	O	546.76
373	CT	185.0	180.0	237	2	O	780.59
374	CT	180.0	168.0	186	2	O	967.74
375	CT	102.5	75.0	165	1	O	621.21
376	CT	161.0	157.7	182	2	O	884.62
377	CT	119.0	114.5	227	1	O	524.23
378	CT	100.0	94.0	166	1	O	602.41
379	CT	150.0	113.0	134	1	O	1 119.40
380	CT	145.0	139.0	191	2	O	759.16
381	CT	209.0	205.0	200	1	N	1 045.00
382	CT	90.0	88.0	127	1	O	708.66
383	CT	200.0	192.0	144	1	O	1 388.89
384	CT	255.0	235.0	155	1	O	1 645.16
385	CT	239.0	209.0	146	1	O	1 636.99
386	CT	239.0	209.0	146	1	O	1 636.99
387	CT	89.0	89.0	80	1	N	1 112.50
388	CT	165.0	160.0	210	2	O	785.71
389	CT	255.0	235.0	205	2	O	1 243.90
390	CT	95.0	93.5	115	1	O	826.09
391	CT	153.0	190.0	175	1	O	874.29
392	CT	92.9	93.4	80	1	N	1 161.25
393	CT	216.0	207.0	248	2	O	870.97
394	CT	170.0	160.0	210	2	O	809.52
395	CT	325.0	321.0	312	1	N	1 041.67
396	CT	125.0	150.0	130	1	O	961.54
397	CT	220.0	200.0	141	1	O	1 560.28
398	CT	210.0	203.0	168	1	O	1 250.00
399	CT	260.0	240.0	113	1	O	2 300.88
400	CT	76.9	74.4	87	1	O	883.91
401	CT	173.5	166.5	103	1	N	1 684.47
402	CT	135.0	131.0	140	2	N	964.29
403	CT	116.3	114.0	110	1	O	1 057.27
404	CT	240.0	210.0	181	2	N	1 325.97
405	CT	175.0	155.0	187	1	O	935.83
406	CT	265.0	226.0	186	1	O	1 424.73

Appendix A

Data Set 2

In 1991 an MBL4-student, mr Bruce Hatcher, for his research project investigated the attitude of consumers towards frozen vegetables. For the sake of completeness the questionnaire is also displayed. The data set in this appendix contains part of the information in the questionnaire completed by 543 respondents. The variables in the questionnaire used in the data set are as follows:

No	=	The respondent number.	
V2	=	Answer to Question 2:	1 = Never 2 = Almost never 3 = Once or twice a month 4 = Most weeks.
V3	=	Answer to Question 3:	1 = No 2 = Yes.
V4	=	Answer to Question 4:	1 = No 2 = Yes.
Know	=	A count between 1 and 10, which was calculated from the answers to Question 5.1 to Question 5.3. A count of 1 means that the respondent has an extremely poor knowledge of frozen vegetables and answered all the questions incorrectly. On the other hand, a count of 10 means that the respondent has an excellent knowledge of frozen vegetables and answered all the questions correctly.	
V7	=	Marital Status:	1 = Married 2 = Single.
V8	=	Number of children.	
V9	=	Do you earn a salary?	1 = No 2 = Work half day 3 = Work full day.
V10a	=	The number of cooked lunches per week.	
V10b	=	The number of cooked dinners per week.	
V10c	=	The number of times per week fresh vegetables are eaten.	

V11 = The total annual income of the family
1 = Up to R30 000
2 = Between R30 000 and R60 000
3 = More than R60 000.

From the semantic differential scale in Question 1 of the questionnaire the following three average attitude scores were calculated for frozen vegetables:

Health1 = The average of questions 1.2, 1.4, 1.8, 1.9 and 1.12. The score is a measure of the respondent's attitude towards the health aspects of frozen vegetables. A low score (1), reflects an extremely negative attitude regarding the health aspects of frozen vegetables. A high score (5), on the other hand, reflects an extremely positive attitude with regard to the health aspects of frozen vegetables.

Conven1 = The average of questions 1.1, 1.3, 1.7 and 1.11. The score is a measure of the respondent's attitude regarding the convenience aspects of frozen vegetabels. A low score (1), reflects an extremely negative attitude with regard to the convenience aspects of frozen vegetables. A high score (5), on the other hand, reflects an extremely positive attitude regarding the convenience aspects of frozen vegetables.

Financ1 = The average of questions 1.6 and 1.10. The score is a measure of the respondent's attitude with regard to the financial aspects of frozen vegetables. A low score (1), reflects an extremely negative attitude regarding the financial aspects of frozen vegetables. A high score (5), on the other hand, reflects an extremely positive attitude regarding the financial aspects of frozen vegetables.

From Question 6 in the questionnaire the following three attitude scores were calculated for fresh vegetables:

Health2 = The average of questions 6.2, 6.4, 6.8, 6.9 and 6.12. The score has the same significance as the variable Health1, but in this instance, for fresh vegetables.

Conven2 = The average of questions 6.1, 6.3, 6.7 and 6.11. The score has the same significance as the variable Conven1, but in this instance, for fresh vegetables.

Financ2 = The average of questions 6.6 and 6.10. The score has the same significance as the variable Financ1, but in this instance, for fresh vegetables.

The following differences were calculated:

Diff1 = Health1 – Health 2:
The difference between the health scores for frozen and fresh vegetables. This score varies between –4 and +4. The difference is positive and close to 4 if a respondent has a very positive attitude towards the health aspects of frozen vegetables and a very negative attitude with regard to the health aspects of fresh vegetables. If a respondent positions fresh and frozen vegetables close together on the five-point scale, the difference is approximately zero. A negative difference means that the respondent feels very positive with regard to the health aspects of fresh vegetables and very negative with regard to the health aspects of frozen vegetables.

Diff2 = Conven1 – Conven2:
The difference between the convenience scores for frozen and fresh vegetables. This score varies between –4 and +4. The difference is positive and close to 4 if a respondent has a very positive attitude towards the convenience aspects of frozen vegetables and a very negative attitude with regard to the convenience aspects of fresh vegetables. If a respondent positions fresh and frozen vegetables close together on the five-point scale, the difference is approximately zero. A negative difference means that the respondent feels very positive with regard to the convenience aspects of fresh vegetables and very negative with regard to the convenience aspects of frozen vegetables.

Diff3 = Financ1 – Financ2:
The difference between the financial aspect scores for frozen and fresh vegetables. This score varies between –4 and +4. The difference is positive and close to 4 if a respondent has a very positive attitude towards the financial aspects of frozen vegetables and a very negative attitude with regard to the financial aspects of fresh vegetables. If a respondent positions fresh and frozen vegetables close together on the five-point scale, the difference is approximately zero. A negative difference means that the respondent feels very positive with regard to the financial aspects of fresh vegetables and very negative with regard to the financial aspects of frozen vegetables.

Note: Missing data is indicated by '–'.

PLEASE READ INSTRUCTIONS FULLY BEFORE WRITING

For office use only

1 2 3 4

INSTRUCTIONS

- The following questions are about **FROZEN VEGETABLES**

Below are 12 blocks. Each block has two descriptions and a five-point scale.
Please show your opinion by circling one number in each block.
Please answer all 12 questions.

EXAMPLE

Inconvenient	1	2	3	4	5	Convenient

5 6

Circling number 4 shows that in your opinion, frozen vegetables are convenient, but they could be even more convenient (in which case you would have circled number 5).

1 In your opinion, **FROZEN** vegetables are:

1	Inconvenient	1	2	3	4	5	Convenient	7
2	Low quality	1	2	3	4	5	High quality	8
3	High wastage	1	2	3	4	5	No wastage	9
4	No vitamins	1	2	3	4	5	Full of vitamins	10
5	Preservatives used	1	2	3	4	5	No preservatives	11
6	Expensive	1	2	3	4	5	Cheap	12
7	Slow to cook	1	2	3	4	5	Quick to cook	13
8	No taste	1	2	3	4	5	Full of taste	14
9	Not healthy	1	2	3	4	5	Healthy	15
10	Luxury	1	2	3	4	5	Necessity	16
11	Seasonal	1	2	3	4	5	Always available	17
12	Not fresh	1	2	3	4	5	Fresh	18

2 How often do you eat frozen vegetables?

Never	Almost never	Once or twice a month	Most weeks

19

3 Are frozen vegetables more expensive than fresh vegetables?

No ☐ Yes ☐

20

4 If frozen vegetables were cheaper than fresh vegetables, would you buy more frozen vegetables?

No ☐ Yes ☐

21

5 In YOUR opinion, during the process of making frozen vegetables, the frozen vegetables are:

1	Cooked before being frozen	True	False	Unsure	22
2	Treated with chemicals	True	False	Unsure	23
3	Of only the highest quality	True	False	Unsure	24

INSTRUCTIONS

- The following questions are about **FRESH VEGETABLES**.

Please show your opinion by circling one number in each block.
Please answer all questions

6 In your opinion, **FRESH** vegetables are:

1	Inconvenient	1	2	3	4	5	Convenient	25
2	Low quality	1	2	3	4	5	High quality	26
3	High wastage	1	2	3	4	5	No wastage	27
4	No vitamins	1	2	3	4	5	Full of vitamins	28
5	Preservatives used	1	2	3	4	5	No preservatives	29
6	Expensive	1	2	3	4	5	Cheap	30
7	Slow to cook	1	2	3	4	5	Quick to cook	31
8	No taste	1	2	3	4	5	Full of taste	32
9	Not healthy	1	2	3	4	5	Healthy	33
10	Luxury	1	2	3	4	5	Necessity	34
11	Seasonal	1	2	3	4	5	Always available	35
12	Not fresh	1	2	3	4	5	Fresh	36

Please, will you supply the following details, by circling the relevant box:

7	Married	Single	37
8	Children: How many		38
9	Do you earn a salary?	No / Halfday / Fullday	39

10	Cooked lunches per week	0	1	2	3	4	5	6	7	40
	Cooked dinners per week	0	1	2	3	4	5	6	7	41
	Eat fresh vegetables per week	0	1	2	3	4	5	6	7	42

11	Total annual income	Up to R30 000	R31 000 - R 60 000	Over R60 000	43

No	V2	V3	V4	Know	V7	V8	V9	V10a	V10b	V10c	V11	Health1	Conven1	Financ1	Health2	Conven2	Financ2	Diff1	Diff2	Diff3
1	4	2	2	10	1	2	1	0	7	3	3	3.20	4.50	4.0	4.20	3.25	4.5	−1.00	1.25	−0.5
2	4	2	2	8	2	3	3	0	7	5	1	4.20	5.00	3.0	4.60	1.75	3.0	−0.40	3.25	0.0
3	4	1	–	9	1	0	1	1	6	2	2	2.80	5.00	4.0	4.00	1.50	3.0	−1.20	3.50	1.0
4	3	2	2	8	2	2	3	5	0	5	2	3.40	4.50	1.5	4.40	2.50	4.0	−1.00	2.00	−2.5
5	4	2	2	9	1	3	3	1	7	7	3	3.60	4.25	2.0	4.20	2.25	3.5	−0.60	2.00	−1.5
6	2	2	2	9	1	0	3	3	5	5	3	3.20	3.75	2.0	4.80	2.75	3.5	−1.60	1.00	−1.5
7	2	2	1	7	1	1	1	0	5	5	3	4.40	4.75	3.0	5.00	2.50	4.5	−0.60	2.25	−1.5
8	3	2	1	4	1	2	1	0	7	7	–	3.00	4.75	1.5	5.00	3.50	5.0	−2.00	1.25	−3.5
9	3	2	2	10	1	0	3	5	7	7	2	4.20	4.25	3.0	4.60	3.00	4.0	−0.40	1.25	−1.0
10	3	2	2	9	2	2	3	5	2	5	2	2.20	4.50	1.0	5.00	2.50	5.0	−2.80	2.00	−4.0
11	4	1	1	4	2	0	3	0	2	1	1	3.60	4.75	3.0	4.60	3.00	3.0	−1.00	1.75	0.0
12	3	2	1	8	1	2	2	1	5	5	3	3.60	3.00	2.0	4.20	2.50	3.0	−0.60	0.50	−1.0
13	4	–	2	6	2	0	3	7	7	2	2	3.00	4.00	2.0	3.80	3.25	3.0	−0.80	0.75	−1.0
14	4	1	–	8	1	1	3	0	7	7	–	5.00	5.00	4.5	4.20	1.50	3.0	0.80	3.50	1.5
15	4	2	2	6	1	3	3	0	6	6	2	3.80	5.00	3.5	5.00	2.75	4.0	−1.20	2.25	−0.5
16	4	2	1	10	1	4	1	1	7	2	–	5.00	5.00	1.0	5.00	3.00	4.0	0.00	2.00	−3.0
17	4	1	–	8	1	2	3	0	7	7	2	5.00	5.00	4.0	3.40	1.33	3.0	1.60	3.67	1.0
18	1	2	2	8	1	2	1	7	7	3	1	4.60	4.00	3.0	5.00	4.00	3.0	−0.40	0.00	0.0
19	1	–	–	5	1	1	1	7	7	7	1	3.33	5.00	1.0	4.60	5.00	4.0	−1.27	0.00	−3.0
20	1	2	‖	4	2	–	3	2	7	3	1	1.20	3.00	1.0	5.00	3.50	5.0	−3.80	−0.50	−4.0
21	4	2	2	5	1	4	2	1	6	5	1	3.20	4.50	2.0	4.00	3.00	3.5	−0.80	1.50	−1.5
22	3	2	2	6	1	3	1	7	7	3	2	3.60	4.50	1.0	4.60	2.25	3.5	−1.00	2.25	−2.5
23	4	2	2	6	2	2	2	2	6	6	1	3.60	4.75	2.0	4.60	2.00	4.0	−1.00	2.75	−2.0
24	4	2	2	9	2	0	3	5	0	1	2	4.00	5.00	4.0	5.00	2.00	1.0	−1.00	3.00	3.0
25	4	2	2	7	1	2	3	0	5	4	3	4.20	4.25	3.0	4.60	2.50	1.5	−0.40	1.75	1.5
26	4	2	2	6	1	2	1	0	7	6	2	3.40	4.75	3.5	5.00	3.75	3.0	−1.60	1.00	0.5
27	4	2	2	6	1	1	1	0	7	5	3	3.80	3.25	4.0	4.80	3.50	3.5	−1.00	−0.25	0.5
28	4	1	–	8	2	–	3	0	7	4	2	5.00	5.00	4.0	3.00	2.00	4.0	2.00	3.00	0.0
29	3	2	2	8	1	1	3	2	3	1	2	2.60	4.50	3.0	3.40	1.25	2.0	−0.80	3.25	1.0
30	1	2	2	6	1	–	1	7	7	6	1	3.20	4.75	2.0	3.60	2.00	3.0	−0.40	2.75	−1.0
31	4	2	1	9	2	0	3	0	7	3	2	3.40	5.00	3.0	4.60	3.00	3.0	−1.20	2.00	0.0
32	4	2	2	8	1	1	1	2	7	3	3	4.00	4.75	1.0	3.00	1.75	3.5	1.00	3.00	−2.5
33	3	2	2	3	2	0	3	0	6	2	2	3.40	4.25	2.5	4.20	2.75	3.5	−0.80	1.50	−1.0
34	4	2	2	8	1	1	3	2	7	7	1	4.60	4.50	3.0	4.00	2.75	3.5	0.60	1.75	−0.5
35	4	2	2	8	1	3	1	4	3	2	3	3.40	4.25	1.5	3.40	2.25	4.0	0.00	2.00	−2.5
36	4	2	2	6	1	1	3	2	7	5	1	3.00	4.75	2.0	4.80	2.00	5.0	−1.80	2.75	−3.0
37	2	2	1	9	1	2	1	5	5	7	3	1.80	4.00	2.0	3.80	3.25	4.0	−2.20	0.75	−2.0
38	4	2	2	2	1	2	3	0	6	7	2	3.40	4.50	3.0	4.20	2.25	4.0	−0.80	2.25	−1.0
39	4	1	2	–	1	0	-	-	-	-	1	1.00	5.00	3.0	3.00	4.00	5.0	−2.00	1.00	−2.0
40	4	2	2	10	1	2	1	1	5	3	2	4.40	4.00	2.5	3.00	1.25	4.0	1.40	2.75	−1.5
41	4	1	1	6	1	1	2	0	7	5	2	5.00	5.00	4.0	3.60	1.25	4.5	1.40	3.75	−0.5
42	4	2	1	9	1	2	3	0	6	6	2	4.20	4.00	2.5	4.00	3.00	4.0	0.20	1.00	−1.5
43	4	2	2	4	1	0	3	0	7	7	1	3.60	5.00	2.5	5.00	3.75	4.5	−1.40	1.25	−2.0
44	4	2	2	10	1	3	2	1	6	7	3	4.00	4.50	2.5	4.20	3.25	3.5	−0.20	1.25	−1.0
45	4	2	2	10	1	3	1	7	7	7	3	4.40	4.50	2.5	3.40	1.75	3.5	1.00	2.75	−1.0
46	2	2	1	9	1	1	1	0	5	5	1	3.60	4.75	2.0	3.80	3.00	4.5	−0.20	1.75	−2.5
47	4	2	1	7	1	0	2	1	5	1	3	3.60	4.75	4.0	5.00	2.50	3.5	−1.40	2.25	0.5
48	4	2	1	9	1	2	2	5	7	7	1	3.60	3.75	4.0	3.60	4.25	4.0	0.00	−0.50	0.0
49	4	1	1	8	1	2	1	0	7	3	2	4.00	4.75	3.0	3.60	2.25	3.5	0.40	2.50	−0.5
50	4	2	2	6	1	2	1	0	7	5	-	3.00	4.50	2.5	4.00	2.50	3.5	−1.00	2.00	−1.0

No	V2	V3	V4	Know	V7	V8	V9	V10a	V10b	V10c	V11	Health1	Conven1	Financ1	Health2	Conven2	Financ2	Diff1	Diff2	Diff3
51	4	–	–	7	1	1	2	1	6	6	2	4.80	5.00	3.0	4.60	3.00	3.0	0.20	2.00	0.0
52	4	2	1	10	1	4	2	0	6	5	3	4.40	4.67	3.5	5.00	3.50	4.0	–0.60	1.17	–0.5
53	4	1	2	9	2	2	3	0	7	2	1	5.00	5.00	5.0	4.20	3.00	3.0	0.80	2.00	2.0
54	4	2	1	–	1	3	1	0	1	7	1	4.00	3.25	4.5	3.67	–	–	0.33	–	–
55	1	1	2	10	2	5	2	0	7	2	1	4.80	5.00	4.0	4.60	4.25	4.0	0.20	0.75	0.0
56	4	2	1	9	1	3	3	0	7	2	3	3.20	4.00	2.5	2.80	2.25	2.5	0.40	1.75	0.0
57	3	2	1	3	1	3	2	3	7	4	3	2.8	4.25	2.5	5.0	2.50	4.0	–2.2	1.75	–1.5
58	3	2	2	8	1	3	1	0	7	7	3	2.6	4.00	1.5	4.0	2.75	4.5	–1.4	1.25	–3.0
59	4	2	2	6	1	4	1	7	0	3	1	4.6	5.00	2.0	4.4	2.00	3.5	0.2	3.00	–1.5
60	4	2	2	4	1	2	1	2	7	6	3	3.6	3.50	4.0	3.6	2.00	2.0	0.0	1.50	2.0
61	4	2	2	6	1	3	3	7	5	4	2	4.0	4.75	3.5	4.0	2.00	3.0	0.0	2.75	0.5
62	4	2	2	6	1	8	1	7	0	7	3	4.2	5.00	3.0	2.2	1.25	3.0	2.0	3.75	0.0
63	2	2	–	6	2	–	1	7	6	7	1	4.6	4.50	4.5	4.8	4.25	4.5	–0.2	0.25	0.0
64	1	2	1	–	1	–	3	–	7	7	2	–	–	–	5.0	4.00	4.0	–	–	–
65	4	2	2	9	1	2	1	1	5	6	2	4.0	4.75	3.0	4.4	4.50	4.5	–0.4	0.25	–1.5
66	4	2	2	8	1	1	2	0	7	3	–	3.0	4.00	1.0	3.0	1.50	3.0	0.0	2.50	–2.0
67	2	2	1	10	1	3	2	1	7	7	1	4.8	4.75	2.0	4.6	3.75	4.5	0.2	1.00	–2.5
68	4	2	1	5	1	2	2	1	6	5	2	3.2	4.50	2.5	4.6	2.75	4.0	–1.4	1.75	–1.5
69	4	2	1	8	1	0	1	1	5	2	3	4.2	5.00	4.0	–	–	–	–	–	–
70	4	1	–	8	1	1	3	0	4	0	3	5.0	5.00	5.0	3.0	1.00	2.5	2.0	4.00	2.5
71	4	2	2	10	2	–	1	2	2	2	1	4.2	4.75	3.0	5.0	3.00	4.0	–0.8	1.75	–1.0
72	4	2	2	10	1	2	2	1	5	4	1	3.6	4.50	3.5	4.2	1.75	2.5	–0.6	2.75	1.0
73	3	2	2	9	1	2	3	0	7	7	1	3.2	4.75	1.0	4.0	3.75	4.0	–0.8	1.00	–3.0
74	4	1	–	10	1	1	3	0	4	1	1	4.0	4.50	3.5	4.2	1.50	3.5	–0.2	3.00	0.0
75	4	2	2	9	1	2	3	0	7	3	2	3.6	5.00	3.5	4.6	1.50	1.0	–1.0	3.50	2.5
76	4	2	2	7	1	2	1	0	6	6	2	3.4	4.75	1.0	4.2	1.75	4.5	–0.8	3.00	–3.5
77	3	2	2	6	1	2	1	3	7	7	1	4.0	4.75	2.0	4.6	3.50	4.0	–0.6	1.25	–2.0
78	3	2	2	8	1	3	2	0	7	6	2	4.0	4.00	1.0	3.8	3.25	4.5	0.2	0.75	–3.5
79	4	1	–	10	2	3	1	5	6	7	2	5.0	5.00	5.0	3.4	2.00	3.5	1.6	3.00	1.5
80	4	1	2	–	1	0	1	0	7	7	–	4.8	5.00	4.0	5.0	5.00	5.0	–0.2	0.00	–1.0
81	4	2	2	9	1	2	2	1	6	4	2	3.2	4.25	2.0	4.6	2.50	4.5	–1.4	1.75	–2.5
82	3	2	1	7	1	2	2	0	7	7	3	3.2	4.50	2.0	3.0	4.75	2.0	0.2	–0.25	0.0
83	4	1	2	10	1	3	1	2	6	5	3	3.6	4.50	3.0	4.8	2.75	4.5	–1.2	1.75	–1.5
84	4	2	2	7	1	2	1	1	5	1	2	5.0	4.00	1.0	5.0	1.75	3.0	0.0	2.25	–2.0
85	3	2	2	4	1	3	3	3	3	7	1	4.6	4.25	1.0	5.0	4.50	5.0	–0.4	–0.25	–4.0
86	4	1	2	10	1	2	1	0	7	7	1	4.2	4.75	3.5	3.6	2.50	3.5	0.6	2.25	0.0
87	4	1	2	–	1	2	1	0	7	3	–	5.0	5.00	5.0	5.0	4.00	5.0	0.0	1.00	0.0
88	2	1	1	4	1	2	1	5	4	7	1	1.6	1.00	1.0	4.0	3.25	4.5	–2.4	–2.25	–3.5
89	3	2	2	6	1	4	2	2	5	6	2	3.0	5.00	2.0	4.4	3.00	4.0	–1.4	2.00	–2.0
90	4	2	1	2	2	0	1	2	7	5	1	3.4	4.25	4.5	4.6	3.50	5.0	–1.2	0.75	–0.5
91	3	1	–	4	1	0	3	1	7	6	3	3.2	4.50	3.0	5.0	3.25	4.5	–1.8	1.25	–1.5
92	4	2	2	9	2	0	3	0	7	7	1	4.6	4.50	2.0	4.2	3.00	4.5	0.4	1.50	–2.5
93	3	2	1	8	1	3	3	5	7	7	2	3.6	4.75	2.0	5.0	3.50	4.5	–1.4	1.25	–2.5
94	4	1	2	8	1	2	1	2	7	6	2	1.0	1.00	2.0	2.6	1.00	5.0	–1.6	0.00	–3.0
95	4	2	2	10	1	3	1	6	0	7	2	3.2	2.75	5.0	4.2	3.25	3.5	–1.0	–0.50	1.5
96	4	2	2	10	1	3	3	2	5	5	2	4.0	4.00	3.0	3.8	2.25	3.5	0.2	1.75	–0.5
97	4	2	2	8	2	1	1	0	6	3	2	4.6	4.75	2.0	–	1.00	3.0	–	3.75	–1.0
98	4	2	2	9	1	3	1	1	7	3	2	3.6	4.50	2.0	4.8	2.50	3.5	–1.2	2.00	–1.5
99	1	2	2	7	2	0	2	3	7	2	2	3.4	4.50	2.0	4.2	2.75	4.5	–0.8	1.75	–2.5
100	2	2	1	2	2	2	3	0	6	6	1	3.0	3.25	4.0	4.6	3.50	4.0	–1.6	–0.25	0.0

No	V2	V3	V4	Know	V7	V8	V9	V10a	V10b	V10c	V11	Health1	Conven1	Financ1	Health2	Conven2	Financ2	Diff1	Diff2	Diff3
101	4	2	1	7	1	1	1	7	7	7	–	4.0	4.00	4.5	5.0	4.00	5.0	–1.0	0.00	–0.5
102	4	2	1	7	1	1	1	7	7	7	–	4.0	4.00	4.5	5.0	4.00	5.0	–1.0	0.00	–0.5
103	4	2	2	9	1	4	1	0	5	3	1	5.0	4.75	4.5	3.6	2.75	4.5	1.4	2.00	0.0
104	3	2	2	9	1	4	1	0	7	7	2	4.2	4.75	1.0	4.8	3.00	4.0	–0.6	1.75	–3.0
105	4	2	1	10	1	0	3	0	5	3	1	4.2	5.00	3.0	5.0	3.50	4.5	–0.8	1.50	–1.5
106	4	2	2	8	1	0	3	1	7	3	1	4.8	5.00	4.0	2.4	1.75	3.5	2.4	3.25	0.5
107	4	2	2	9	1	0	1	6	0	6	1	3.6	4.50	3.0	4.2	3.25	4.0	–0.6	1.25	–1.0
108	4	2	2	9	1	3	1	1	4	2	1	5.0	5.00	1.0	4.2	1.25	3.0	0.8	3.75	–2.0
109	4	2	2	7	1	1	1	0	7	3	2	4.2	5.00	3.0	4.4	2.75	3.5	–0.2	2.25	–0.5
110	–	1	–	8	1	–	3	0	7	2	3	–	–	–	–	1.00	–	–	–	–
111	4	2	2	7	1	2	2	3	5	4	2	4.0	5.00	1.0	3.6	5.00	1.0	0.4	0.00	0.0
112	4	1	–	9	1	0	3	0	3	2	1	3.4	5.00	3.0	4.2	3.50	4.5	–0.8	1.50	–1.5
113	4	2	1	10	1	2	–	0	7	7	–	4.20	5.00	4.0	4.4	5.00	4.5	–0.20	0.00	–0.5
114	4	2	2	8	1	2	2	5	7	7	2	4.60	5.00	2.5	4.0	2.50	5.0	0.60	2.50	–2.5
115	4	2	2	10	1	2	3	0	7	3	1	4.80	4.75	4.0	4.2	1.75	5.0	0.40	3.00	–1.0
116	3	2	–	9	1	0	1	0	6	5	2	3.60	4.50	1.0	4.0	2.75	4.5	–0.40	1.75	–3.5
117	4	2	2	9	1	4	3	0	6	3	1	4.00	5.00	5.0	4.4	3.00	3.5	–0.40	2.00	1.5
118	4	2	2	10	1	0	3	0	7	7	1	5.00	5.00	3.0	3.0	1.50	3.0	2.00	3.50	0.0
119	2	2	1	–	1	0	–	–	–	7	–	–	5.00	–	5.0	–	–	–	–	–
120	4	2	2	7	1	2	1	0	5	5	3	3.40	2.50	4.5	2.0	4.00	4.5	1.40	–1.50	0.0
121	4	2	2	10	1	4	1	7	0	7	1	5.00	4.75	3.5	3.8	2.75	4.0	1.20	2.00	–0.5
122	1	2	2	6	1	4	1	4	7	3	1	2.75	3.50	1.0	4.2	3.00	4.0	–1.45	0.50	–3.0
123	4	2	–	9	1	3	2	0	7	3	–	3.80	4.75	1.5	4.4	1.25	2.0	–0.60	3.50	–0.5
124	4	–	1	5	1	3	3	0	7	3	1	3.80	5.00	3.5	4.6	3.50	4.0	–0.80	1.50	–0.5
125	4	2	2	10	1	2	3	0	7	3	2	4.80	5.00	1.0	5.0	1.50	4.0	–0.20	3.50	–3.0
126	4	2	2	9	1	4	1	7	7	7	1	4.20	5.00	3.0	5.0	4.00	3.0	–0.80	1.00	0.0
127	3	2	1	7	1	6	1	3	7	4	–	1.40	4.75	1.5	4.8	2.50	4.0	–3.40	2.25	–2.5
128	4	2	1	5	1	1	1	4	1	3	2	3.40	5.00	2.0	4.4	2.50	4.5	–1.00	2.50	–2.5
129	4	2	2	2	1	3	1	1	6	6	–	3.80	5.00	1.5	4.2	3.75	4.0	–0.40	1.25	–2.5
130	4	2	1	8	1	1	2	0	7	7	2	4.20	5.00	4.0	4.2	2.75	4.5	0.00	2.25	–0.5
131	4	2	2	6	1	3	1	1	7	1	3	3.40	4.50	1.0	4.2	1.00	2.0	–0.80	3.50	–1.0
132	4	2	2	6	1	3	1	3	7	3	1	3.60	4.75	1.0	3.0	4.00	3.0	0.60	0.75	–2.0
133	4	2	2	7	1	3	1	3	6	6	2	4.80	4.75	1.5	4.6	2.50	4.5	0.20	2.25	–3.0
134	2	2	2	6	1	2	1	0	7	7	2	3.20	4.75	1.0	4.2	3.25	4.0	–1.00	1.50	–3.0
135	4	2	2	10	1	2	1	0	7	7	2	5.00	4.75	2.5	4.6	3.50	4.0	0.40	1.25	–1.5
136	3	2	2	10	1	3	3	0	6	4	2	4.60	4.50	1.0	2.8	1.00	4.0	1.80	3.50	–3.0
137	2	2	2	10	1	1	1	0	7	7	2	4.20	4.75	1.0	4.2	2.75	4.5	0.00	2.00	–3.5
138	4	2	2	8	1	0	1	1	5	5	2	4.80	4.75	3.0	4.6	3.50	4.5	0.20	1.25	–1.5
139	3	2	1	8	1	0	1	2	5	7	–	3.20	5.00	2.5	3.2	3.50	3.0	0.00	1.50	–0.5
140	4	2	2	9	1	3	2	1	5	6	1	4.20	4.50	2.5	4.8	2.50	3.5	–0.60	2.00	–1.0
141	3	2	2	10	1	–	1	2	7	7	2	3.80	5.00	1.0	4.2	1.75	4.0	–0.40	3.25	–3.0
142	4	1	–	6	1	–	1	7	0	7	–	4.80	4.50	4.0	4.0	3.25	4.0	0.80	1.25	0.0
143	4	2	2	8	1	–	2	2	–	–	1	3.00	5.00	1.0	5.0	3.25	5.0	–2.00	1.75	–4.0
144	4	2	2	8	1	2	2	1	7	4	2	4.80	4.75	1.0	5.0	2.25	4.5	–0.20	2.50	–3.5
145	4	2	2	10	1	1	3	0	7	6	2	3.40	4.50	3.0	4.0	2.25	3.5	–0.60	2.25	–0.5
146	3	2	1	10	1	1	1	7	0	5	1	3.80	4.00	4.0	5.0	2.00	4.0	–1.20	2.00	0.0
147	4	2	1	9	2	–	3	3	6	6	1	3.80	5.00	3.0	4.8	3.50	4.0	–1.00	1.50	–1.0
148	4	2	2	8	1	3	1	0	7	6	1	5.00	5.00	4.5	4.6	3.75	5.0	0.40	1.25	–0.5
149	4	2	2	10	1	–	1	2	5	5	1	4.40	4.50	4.0	4.2	2.50	4.0	0.20	2.00	0.0
150	4	2	2	10	1	0	1	0	7	7	1	4.60	5.00	4.5	4.8	3.75	5.0	–0.20	1.25	–0.5

No	V2	V3	V4	Know	V7	V8	V9	V10a	V10b	V10c	V11	Health1	Conven1	Financ1	Health2	Conven2	Financ2	Diff1	Diff2	Diff3
151	4	2	2	9	1	3	1	2	7	6	2	3.80	3.75	2.0	3.0	2.00	3.0	0.80	1.75	−1.0
152	4	2	2	10	2	–	3	0	7	7	2	4.60	4.50	3.0	4.0	3.00	3.0	0.60	1.50	0.0
153	3	2	1	9	1	–	1	7	0	4	2	4.20	4.75	4.0	3.8	3.00	3.0	0.40	1.75	1.0
154	4	2	2	10	1	2	1	2	5	6	2	4.00	5.00	2.5	4.8	3.00	4.0	−0.80	2.00	−1.5
155	3	2	1	10	1	1	1	3	6	6	2	3.80	4.25	4.5	4.6	3.75	4.0	−0.80	0.50	0.5
156	3	2	2	10	1	3	1	4	2	7	3	1.00	4.50	1.0	4.8	4.25	5.0	−3.80	0.25	−4.0
157	–	2	1	7	1	2	1	2	7	6	1	3.20	4.25	2.5	4.2	3.00	4.0	−1.00	1.25	−1.5
158	2	1	1	9	1	2	1	1	7	7	3	1.20	4.75	2.0	5.0	2.50	2.5	−3.80	2.25	−0.5
159	–	2	2	9	1	2	1	4	0	4	1	5.00	4.75	1.0	4.4	3.00	4.0	0.60	1.75	−3.0
160	4	2	2	9	1	3	3	0	7	7	3	3.20	4.75	2.0	4.4	4.00	4.0	−1.20	0.75	−2.0
161	3	2	2	8	1	1	1	1	5	6	3	3.60	4.00	3.5	4.4	3.50	3.5	−0.80	0.50	0.0
162	–	2	1	7	1	0	3	0	7	7	–	3.00	4.75	3.0	5.0	2.75	2.5	−2.00	2.00	0.5
163	4	1	2	8	2	–	1	2	7	7	1	3.80	5.00	4.0	4.0	3.25	4.0	−0.20	1.75	0.0
164	4	2	2	8	1	–	1	0	6	4	2	4.40	5.00	3.5	4.6	2.25	3.5	−0.20	2.75	0.0
165	4	2	2	9	1	4	2	3	6	3	2	4.00	4.67	5.0	–	5.00	5.0	–	−0.33	0.0
166	4	2	2	2	1	2	1	0	5	5	2	2.60	3.00	2.5	3.8	3.25	2.0	−1.20	−0.25	0.5
167	4	2	2	9	1	0	3	1	6	1	1	3.60	3.75	1.5	4.2	2.00	4.0	−0.60	1.75	−2.5
168	4	2	2	10	1	2	3	0	7	7	2	4.20	5.00	3.0	4.4	3.50	4.0	−0.20	1.50	−1.0
169	4	2	2	5	1	0	3	0	7	7	1	3.80	4.75	3.0	3.8	2.50	4.0	0.20	2.25	−1.0
170	4	2	2	2	1	–	1	0	7	7	1	3.40	4.25	3.5	4.0	3.25	3.5	−0.60	1.00	0.0
171	4	2	2	7	1	3	1	1	6	3	3	4.20	5.00	2.0	3.4	1.50	3.0	0.80	3.50	−1.0
172	3	1	2	9	1	3	1	0	6	7	3	3.20	4.25	1.5	4.4	2.75	5.0	−1.20	1.50	−3.5
173	2	2	2	7	1	2	1	0	7	7	3	2.60	4.25	1.5	4.6	3.00	3.5	−2.00	1.25	−2.0
174	3	2	2	9	–	–	–	–	–	–	–	–	–	–	4.8	5.00	4.0	–	–	–
175	2	2	1	8	1	0	1	1	7	7	2	2.75	4.00	4.5	1.0	2.25	1.0	1.75	1.75	3.5
176	4	2	1	2	1	2	1	1	7	4	3	3.60	4.75	3.5	3.4	2.75	4.0	0.20	2.00	−0.5
177	4	2	2	6	1	1	1	2	5	4	3	3.60	3.50	3.5	3.6	2.67	4.0	0.00	0.83	−0.5
178	3	2	2	10	1	5	1	1	7	7	3	2.60	4.50	2.0	5.0	2.50	5.0	−2.40	2.00	−3.0
179	4	2	2	10	1	2	1	0	7	7	1	5.00	5.00	–	3.0	5.00	–	2.00	0.00	–
180	4	2	2	6	1	0	3	2	7	7	2	3.20	3.50	2.0	4.2	2.25	4.0	−1.00	1.25	−2.0
181	4	2	2	9	–	–	1	0	5	4	–	4.80	5.00	3.0	2.8	3.50	1.5	2.00	1.50	1.5
182	4	2	2	5	1	3	1	0	7	7	3	3.20	4.25	3.0	4.2	2.25	4.5	−1.00	2.00	−1.5
183	4	'	–	6	1	2	3	0	7	7	1	3.80	5.00	2.5	4.4	2.25	3.0	−0.60	2.75	−0.5
184	2	2	1	6	1	1	2	0	7	7	1	3.60	5.00	1.0	4.2	2.75	5.0	−0.60	2.25	−4.0
185	4	2	2	8	1	3	1	0	7	2	2	3.60	5.00	1.0	3.0	2.25	2.0	0.60	2.75	−1.0
186	3	2	1	6	1	1	2	1	6	3	1	4.20	4.50	4.0	4.6	3.75	4.0	−0.40	0.75	0.0
187	4	2	1	9	2	0	1	0	7	7	1	4.00	5.00	2.0	4.8	2.75	4.0	−0.80	2.25	−2.0
188	4	2	2	5	1	5	3	0	7	3	3	3.20	3.75	3.0	4.4	2.00	3.5	−1.20	1.75	−0.5
189	2	2	2	8	1	2	1	3	7	6	3	4.20	4.50	3.0	5.0	4.50	5.0	−0.80	0.00	−2.0
190	3	2	1	5	2	–	2	2	7	7	2	4.40	5.00	1.0	3.4	1.50	4.0	1.00	3.50	−3.0
191	4	2	1	9	1	1	3	4	7	0	3	5.00	4.75	3.0	1.6	2.25	4.0	3.40	2.50	−1.0
192	4	2	2	7	2	2	3	2	7	7	1	4.00	5.00	4.0	4.4	1.25	3.5	−0.40	3.75	0.5
193	2	2	1	10	1	–	3	2	5	6	1	3.00	4.75	3.0	3.6	3.25	3.5	−0.60	1.50	−0.5
194	4	2	2	2	1	6	3	2	3	2	2	4.80	4.75	3.0	2.6	1.75	2.5	2.20	3.00	0.5
195	4	1	2	3	1	2	3	1	4	4	1	3.40	4.75	3.5	4.0	3.75	5.0	−0.60	1.00	−1.5
196	4	2	1	8	1	3	2	0	6	6	3	3.20	5.00	3.0	4.6	2.75	4.5	−1.40	2.25	−1.5
197	4	1	2	7	1	–	1	0	7	2	1	4.80	5.00	4.0	2.6	1.50	1.0	2.20	3.50	3.0
198	4	1	–	8	1	–	3	0	7	7	2	5.00	5.00	4.0	4.8	1.00	4.0	0.20	4.00	0.0
199	1	1	1	6	1	1	1	0	7	7	1	2.40	5.00	5.0	5.0	1.50	4.5	−2.60	3.50	0.5
200	4	2	–	10	1	2	1	6	6	6	2	4.00	5.00	4.5	5.0	3.67	5.0	−1.00	1.33	−0.5

No	V2	V3	V4	Know	V7	V8	V9	V10a	V10b	V10c	V11	Health1	Conven1	Financ1	Health2	Conven2	Financ2	Diff1	Diff2	Diff3
201	4	1	2	10	1	4	3	0	7	5	2	4.80	4.75	3.0	4.0	1.00	3.0	0.80	3.75	0.0
202	4	2	2	8	1	3	1	2	7	7	2	5.00	4.75	1.5	5.0	5.00	2.5	0.00	–0.25	–1.0
203	4	2	1	7	1	0	1	2	7	5	1	3.20	3.50	2.0	5.0	4.00	4.0	–1.80	–0.50	–2.0
204	4	1	–	9	2	2	1	1	7	7	3	5.00	5.00	5.0	4.0	2.00	3.0	1.00	3.00	2.0
205	2	2	2	3	1	1	1	4	5	7	1	3.60	3.25	3.5	3.8	3.00	2.5	–0.20	0.25	1.0
206	2	2	1	10	1	1	3	2	7	7	3	3.40	4.00	1.0	5.0	4.00	5.0	–1.60	0.00	–4.0
207	4	2	2	9	1	2	2	6	7	7	1	4.40	5.00	3.0	3.6	1.25	3.5	0.80	3.75	–0.5
208	4	2	1	7	2	–	1	3	6	7	–	3.80	4.25	2.5	5.0	3.00	2.5	–1.20	1.25	0.0
209	2	–	2	7	1	0	1	0	7	7	2	1.20	3.75	1.5	5.0	3.75	3.5	–3.80	0.00	–2.0
210	4	2	2	5	1	2	1	1	5	6	2	3.40	4.75	2.0	4.6	3.00	4.5	–1.20	1.75	–2.5
211	4	1	–	9	2	–	3	0	7	7	1	3.40	3.75	3.0	4.4	4.00	4.5	–1.00	–0.25	–1.5
212	3	1	–	9	1	4	2	6	4	5	2	4.00	2.75	4.0	3.6	3.25	3.0	0.40	–0.50	1.0
213	4	2	2	8	1	4	1	3	7	7	–	5.00	5.00	4.5	5.0	2.50	5.0	0.00	2.50	–0.5
214	3	2	2	10	1	2	1	0	5	3	2	3.80	5.00	2.0	4.6	3.00	4.5	–0.80	2.00	–2.5
215	4	2	2	9	1	1	2	0	7	2	1	4.80	5.00	2.0	4.6	2.67	5.0	0.20	2.33	–3.0
216	4	2	1	8	2	–	3	0	2	6	1	3.00	5.00	3.0	4.8	2.00	3.5	–1.80	3.00	–0.5
217	4	2	1	8	1	–	1	0	7	5	3	3.40	5.00	3.0	4.4	2.25	3.5	–1.00	2.75	–0.5
218	3	2	2	10	1	2	1	0	7	7	2	4.60	4.75	2.5	4.2	2.50	4.0	0.40	2.25	–1.5
219	4	2	2	8	1	2	3	–	–	–	2	4.40	4.75	1.0	4.2	1.50	5.0	0.20	–3.25	–4.0
220	4	1	–	9	1	–	3	2	5	5	2	4.00	4.50	3.5	4.4	2.25	2.5	–0.40	2.25	1.0
221	4	2	1	6	1	3	2	0	7	7	3	2.60	4.75	2.0	4.8	3.50	4.5	–2.20	1.25	–2.5
222	4	2	1	8	1	3	3	4	7	7	3	3.60	4.25	1.5	4.2	1.25	5.0	–0.60	3.00	–3.5
223	4	2	1	7	1	2	1	1	7	7	2	3.40	4.25	2.5	4.2	2.50	4.0	–0.80	1.75	–1.5
224	4	1	2	4	1	2	3	5	4	6	1	4.60	5.00	4.5	5.0	4.00	4.0	–0.40	1.00	0.5
225	3	2	1	6	1	3	1	4	7	7	3	3.80	5.00	4.0	4.40	4.00	3.5	–0.60	1.00	0.5
226	3	2	2	8	1	1	3	1	5	7	1	3.20	5.00	1.0	4.60	4.25	4.5	–1.40	0.75	–3.5
227	4	2	2	8	2	–	1	6	7	7	1	4.40	4.00	2.0	4.40	3.25	3.0	0.00	0.75	–1.0
228	4	1	2	7	2	0	1	2	6	7	2	4.33	5.00	4.0	3.00	1.67	3.0	1.33	3.33	1.0
229	4	2	2	10	1	2	1	0	7	5	3	4.40	5.00	2.5	4.00	3.25	4.0	0.40	1.75	–1.5
230	4	2	2	9	1	5	3	7	0	7	2	4.60	4.50	1.0	5.00	1.25	5.0	–0.40	3.25	–4.0
231	4	2	2	7	1	2	3	0	5	3	3	3.80	4.50	3.0	4.40	2.25	4.0	–0.60	2.25	–1.0
232	4	2	2	10	1	0	3	2	5	2	3	4.80	5.00	4.0	4.60	2.50	4.5	0.20	2.50	–0.5
233	3	2	1	8	1	3	3	2	5	4	1	2.20	3.75	4.0	5.00	2.75	4.5	–2.80	1.00	–0.5
234	4	2	1	10	2	2	3	0	5	7	2	3.60	4.50	3.5	4.60	3.00	4.0	–1.00	1.50	–0.5
235	3	2	2	8	1	1	3	1	6	5	3	4.80	4.75	1.5	4.80	3.25	4.5	0.00	1.50	–3.0
236	4	2	2	10	1	0	1	–	7	7	2	4.40	4.25	1.0	4.80	3.50	4.0	–0.40	0.75	–3.0
237	4	2	2	10	1	3	3	2	6	–	–	5.00	–	–	–	1.00	1.0	–	–	–
238	4	2	2	2	1	2	–	7	7	1	3	4.80	5.00	3.5	–	1.00	–	–	4.00	–
239	3	2	2	3	2	0	3	0	7	5	1	3.60	4.00	2.0	4.80	2.25	3.5	–1.20	1.75	–1.5
240	3	2	1	10	1	0	1	3	6	7	2	5.00	5.00	1.0	4.20	4.50	4.0	0.80	0.50	–3.0
241	4	1	2	9	1	2	2	1	6	0	3	3.80	4.50	2.5	4.00	3.00	3.5	–0.20	1.50	–1.0
242	4	2	2	6	1	1	3	0	7	5	2	3.20	4.50	2.0	5.00	2.50	5.0	–1.80	2.00	–3.0
243	4	1	2	8	1	1	2	0	7	2	1	5.00	5.00	5.0	3.40	1.00	2.5	1.60	4.00	2.5
244	3	2	1	2	2	–	1	0	6	4	1	3.20	3.25	5.0	3.80	4.25	4.0	–0.60	–1.00	1.0
245	4	2	2	6	1	2	2	0	7	3	3	2.80	4.25	2.0	5.00	2.25	3.5	–2.20	2.00	–1.5
246	4	1	–	8	1	–	1	2	6	3	1	3.00	–	3.0	2.25	4.50	3.5	0.75	–	–0.5
247	2	2	2	6	1	1	3	0	6	4	1	2.60	4.50	3.0	4.40	2.75	4.0	–1.80	1.75	–1.0
248	4	2	2	9	1	0	1	7	7	7	2	3.00	3.75	1.5	4.60	2.75	4.0	–1.60	1.00	–2.5
249	4	2	2	8	1	2	2	1	6	7	2	4.80	5.00	4.0	4.60	4.00	5.0	0.20	1.00	–1.0
250	4	2	1	2	1	1	1	0	7	3	-	1.00	5.00	3.0	5.00	5.00	5.0	–4.00	0.00	–2.0

No	V2	V3	V4	Know	V7	V8	V9	V10a	V10b	V10c	V11	Health1	Conven1	Financ1	Health2	Conven2	Financ2	Diff1	Diff2	Diff3
251	4	1	2	7	1	2	2	1	7	3	1	4.80	4.00	2.0	4.20	3.00	5.0	0.60	1.00	–3.0
252	4	2	2	6	1	–	1	0	7	7	–	3.40	4.50	1.0	3.20	2.00	4.5	0.20	2.50	–3.5
253	4	2	2	7	1	2	1	3	7	3	1	1.20	2.00	2.5	1.60	3.25	3.5	–0.40	–1.25	–1.0
254	3	2	2	9	1	–	1	7	7	7	1	5.00	4.75	3.5	4.60	2.25	3.5	0.40	2.50	0.0
255	4	1	1	8	1	–	1	0	6	7	3	5.00	4.00	–	5.00	–	–	0.00	–	–
256	4	2	2	8	1	–	1	7	5	3	2	5.00	5.00	3.0	5.00	1.75	4.5	0.00	3.25	–1.5
257	4	2	1	8	1	5	1	0	7	7	2	4.40	4.75	1.0	4.60	2.75	4.5	–0.20	2.00	–3.5
258	2	2	1	8	1	–	3	2	6	7	2	1.40	3.75	1.5	5.00	4.00	4.5	–3.60	–0.25	–3.0
259	2	2	1	10	1	1	1	0	7	7	–	3.00	4.75	2.0	5.00	3.25	5.0	–2.00	1.50	–3.0
260	1	2	2	6	2	–	3	2	0	3	1	3.00	3.25	4.0	3.00	3.75	3.0	0.00	–0.50	1.0
261	4	2	2	9	1	2	1	0	7	7	3	2.80	4.50	2.0	4.60	1.50	3.5	–1.80	3.00	–1.5
262	4	–	–	6	1	3	3	1	6	6	2	5.00	5.00	4.5	4.00	1.75	3.0	1.00	3.25	1.5
263	3	2	2	6	1	2	3	0	7	4	2	3.00	4.75	2.0	4.20	2.25	3.0	–1.20	2.50	–1.0
264	4	1	–	6	1	3	1	3	7	2	2	4.20	4.50	4.0	3.80	1.50	3.0	0.40	3.00	1.0
265	4	2	2	5	1	0	3	0	3	3	2	3.60	4.75	4.0	5.00	3.50	5.0	–1.40	1.25	–1.0
266	4	2	2	9	1	0	3	1	5	7	1	5.00	5.00	2.0	4.00	1.75	4.0	1.00	3.25	–2.0
267	3	2	2	7	1	4	3	2	7	1	1	4.00	4.25	4.0	3.80	3.00	3.0	0.20	1.25	1.0
268	4	1	–	1	1	2	3	0	7	0	2	3.80	5.00	2.0	4.00	1.00	3.0	–0.20	4.00	–1.0
269	4	1	–	8	1	4	1	2	7	2	3	4.60	5.00	4.5	4.00	1.50	3.5	0.60	3.50	1.0
270	4	1	–	10	1	0	3	1	7	2	1	4.75	5.00	3.0	4.00	1.00	3.0	0.75	4.00	0.0
271	4	1	–	8	1	0	1	7	1	2	1	4.40	5.00	3.5	2.20	1.00	2.5	2.20	4.00	1.0
272	4	2	2	8	1	2	3	0	7	5	2	5.00	5.00	5.0	5.00	2.75	3.5	0.00	2.25	1.5
273	4	2	2	8	1	–	3	0	4	4	2	5.00	5.00	5.0	5.00	2.75	3.5	0.00	2.25	1.5
274	3	1	–	9	1	1	3	2	7	3	–	3.40	4.50	4.0	5.00	3.50	4.0	–1.60	1.00	0.0
275	4	1	–	6	2	–	3	1	7	2	1	4.80	5.00	3.0	3.80	1.00	3.0	1.00	4.00	0.0
276	4	2	2	10	2	–	1	2	6	5	2	4.20	5.00	2.0	4.40	2.00	3.0	–0.20	3.00	–1.0
277	4	2	2	5	1	–	3	7	7	7	1	5.00	5.00	3.0	5.00	3.00	5.0	0.00	2.00	–2.0
278	4	2	2	8	1	–	2	1	5	2	1	4.00	5.00	2.0	5.00	2.75	3.5	–1.00	2.25	–1.5
279	4	2	1	1	–	–	–	0	5	5	1	2.80	5.00	2.0	2.40	1.50	3.0	0.40	3.50	–1.0
280	4	2	–	9	1	4	2	0	7	4	2	3.60	5.00	4.5	4.60	2.75	5.0	–1.00	2.25	–0.5
281	4	2	1	8	2	–	1	1	6	6	1	5.00	5.00	2.5	3.75	2.00	5.0	1.25	3.00	–2.5
282	3	2	2	10	1	–	2	0	7	6	1	3.60	4.75	3.5	5.00	3.33	3.0	–1.40	1.42	0.5
283	4	1	1	8	1	2	2	2	5	7	3	3.80	4.25	3.5	4.60	3.25	4.0	–0.80	1.00	–0.5
284	2	2	2	9	2	3	2	1	7	6	3	3.40	4.75	1.5	5.00	3.75	5.0	–1.60	1.00	–3.5
285	4	2	–	6	1	0	3	0	7	6	1	2.40	5.00	1.0	5.00	1.75	5.0	–2.60	3.25	–4.0
286	4	2	2	10	2	–	1	3	3	3	1	4.40	5.00	3.0	2.20	1.00	1.0	2.20	4.00	2.0
287	4	2	2	10	1	0	1	1	6	7	1	5.00	5.00	1.0	4.60	1.00	3.0	0.40	4.00	–2.0
288	4	2	2	8	1	0	1	2	7	7	1	4.60	5.00	2.0	4.80	4.00	4.0	–0.20	1.00	–2.0
289	4	2	–	6	1	0	1	7	7	7	2	4.40	4.00	5.0	4.60	4.00	4.0	–0.20	0.00	1.0
290	4	2	2	9	1	1	3	2	5	4	2	4.40	5.00	3.0	4.20	3.00	4.5	0.20	2.00	–1.5
291	4	2	2	9	1	3	2	3	6	6	2	3.60	4.75	2.0	4.00	2.50	3.0	–0.40	2.25	–1.0
292	4	1	–	9	1	–	1	2	6	2	1	4.60	5.00	3.0	2.80	2.00	5.0	1.80	3.00	–2.0
293	3	2	1	6	1	1	1	0	7	4	1	4.60	5.00	3.0	5.00	2.75	4.0	–0.40	2.25	–1.0
294	1	2	2	9	1	0	1	1	5	7	2	4.40	5.00	1.0	5.00	3.75	3.0	–0.60	1.25	–2.0
295	4	2	2	8	1	–	3	0	7	7	1	3.00	4.25	1.0	4.60	3.50	5.0	–1.60	0.75	–4.0
296	4	1	2	10	1	0	3	0	7	4	1	4.60	5.00	4.0	2.80	2.00	3.0	1.80	3.00	1.0
297	4	2	2	10	1	4	3	0	5	7	2	3.60	4.75	1.0	3.80	2.00	5.0	–0.20	2.75	–4.0
298	3	2	2	8	1	3	3	1	4	4	1	3.60	4.50	2.5	4.20	2.50	3.5	–0.60	2.00	–1.0
299	2	2	2	9	1	1	1	0	7	7	1	3.00	4.00	2.5	4.00	3.00	4.0	–1.00	1.00	–1.5
300	4	1	2	9	1	–	1	1	5	4	1	4.40	4.00	4.0	3.75	3.50	3.0	0.65	0.50	1.0

No	V2	V3	V4	Know	V7	V8	V9	V10a	V10b	V10c	V11	Health1	Conven1	Financ1	Health2	Conven2	Financ2	Diff1	Diff2	Diff3
301	4	2	2	6	1	0	1	0	6	7	2	4.20	4.50	3.0	4.80	2.75	5.0	−0.60	1.75	−2.0
302	4	2	2	10	1	2	3	1	6	4	1	4.20	5.00	1.5	4.40	4.50	3.0	−0.20	0.50	−1.5
303	3	2	2	9	2	–	3	1	3	2	2	4.40	4.75	1.0	4.60	3.25	3.5	−0.20	1.50	−2.5
304	2	2	2	10	1	1	1	0	7	7	1	4.20	4.50	2.5	4.60	3.75	4.5	−0.40	0.75	−2.0
305	1	2	1	10	1	2	1	7	7	7	1	2.60	2.00	3.0	2.33	4.00	5.0	0.27	−2.00	−2.0
306	–	2	–	9	1	2	2	0	7	7	2	3.40	4.75	2.5	5.00	3.25	5.0	−1.60	1.50	−2.5
307	2	2	2	6	1	2	1	0	6	6	–	3.40	3.33	1.0	4.67	4.67	4.0	−1.27	−1.34	−3.0
308	2	2	1	9	1	2	2	0	3	7	1	3.40	5.00	1.0	3.20	2.00	3.5	0.20	3.00	−2.5
309	4	2	2	6	1	0	1	6	5	5	1	4.00	4.75	4.5	3.60	3.25	4.0	0.40	1.50	0.5
310	3	2	1	9	1	2	1	3	1	3	3	2.75	3.25	4.5	3.60	3.00	4.0	−0.85	0.25	0.5
311	4	2	2	10	2	–	1	1	4	6	1	4.50	5.00	2.5	4.00	2.75	4.5	0.50	2.25	−2.0
312	3	2	–	7	1	6	2	0	3	3	1	3.60	4.25	1.0	4.20	2.50	3.0	−0.60	1.75	−2.0
313	4	1	–	6	1	2	1	2	7	7	3	3.80	4.25	3.0	4.20	2.50	3.5	−0.40	1.75	−0.5
314	4	2	1	10	2	2	3	0	6	6	1	4.00	5.00	2.5	4.60	2.00	4.5	−0.60	3.00	−2.0
315	4	2	1	7	1	0	2	1	7	6	2	3.20	5.00	3.0	5.00	2.50	4.5	−1.80	2.50	−1.5
316	4	2	1	9	1	1	1	0	6	4	3	4.20	5.00	3.5	4.20	2.50	3.0	0.00	2.50	0.5
317	4	1	2	10	1	2	1	6	1	2	2	5.00	5.00	3.5	5.00	5.00	4.5	0.00	0.00	−1.0
318	3	2	2	8	2	–	1	0	7	7	2	2.80	4.50	1.0	4.40	2.25	3.5	−1.60	2.25	−2.5
319	3	2	1	7	1	–	1	0	7	4	3	2.60	4.75	3.0	4.40	3.00	3.0	−1.80	1.75	0.0
320	4	2	1	4	1	2	1	7	7	7	1	1.00	–	1.0	5.00	5.00	5.0	−4.00	–	−4.0
321	2	2	2	9	2	1	3	0	4	3	2	4.00	4.75	2.5	4.20	3.75	4.5	−0.20	1.00	−2.0
322	4	2	2	8	1	4	1	0	7	7	1	4.00	4.75	2.0	4.00	5.00	1.0	0.00	−0.25	1.0
323	3	2	2	9	1	2	3	2	6	4	3	3.40	4.50	2.5	3.20	3.00	3.0	0.20	1.50	−0.5
324	4	2	2	8	1	2	1	2	2	2	2	3.40	4.75	3.0	4.00	2.50	4.5	−0.60	2.25	−1.5
325	4	2	2	6	1	3	1	2	7	7	3	3.40	4.75	2.5	3.80	2.75	4.5	−0.40	2.00	−2.0
326	4	2	2	9	1	3	2	3	2	6	1	3.80	2.75	2.5	4.20	1.75	3.0	−0.40	1.00	−0.5
327	4	2	2	8	1	–	2	3	3	3	1	4.80	4.50	3.5	3.60	3.25	3.5	1.20	1.25	0.0
328	3	2	2	8	1	2	1	7	7	7	2	4.00	4.25	2.0	4.60	2.75	3.0	−0.60	1.50	−1.0
329	3	2	1	6	1	–	1	7	0	7	3	3.40	5.00	3.0	5.00	4.00	3.0	−1.60	1.00	0.0
330	4	1	–	9	1	5	1	0	6	1	–	3.80	4.00	4.0	4.60	4.00	5.0	−0.80	0.00	−1.0
331	3	1	–	6	1	–	3	6	6	3	2	3.60	4.00	5.0	4.20	3.50	3.5	−0.60	0.50	1.5
332	4	2	1	6	1	–	1	5	7	6	1	3.80	5.00	4.0	4.40	3.25	2.5	−0.60	1.75	1.5
333	4	2	2	9	1	4	1	0	7	1	1	5.00	4.50	2.0	2.80	1.50	3.0	2.20	3.00	−1.0
334	2	2	2	10	1	–	2	7	1	5	2	4.00	2.50	4.5	3.00	2.75	4.0	1.00	−0.25	0.5
335	4	2	1	10	1	2	1	1	7	7	2	3.80	4.75	4.0	4.60	3.75	3.0	−0.80	1.00	1.0
336	3	2	1	7	2	–	3	0	3	3	2	3.40	5.00	3.0	4.00	2.50	3.5	−0.60	2.50	−0.5
337	2	2	2	9	1	3	1	0	5	7	2	3.4	5.00	1.5	4.6	2.75	5.0	−1.2	2.25	−3.5
338	4	1	2	8	1	2	3	0	7	4	2	5.0	4.75	4.5	4.0	3.50	4.0	1.0	1.25	0.5
339	4	2	2	10	1	3	1	0	7	7	2	5.0	5.00	3.0	5.0	2.50	5.0	0.0	2.50	−2.0
340	4	2	2	9	1	–	3	0	7	6	1	3.4	4.75	3.5	5.0	2.25	3.5	−1.6	2.50	0.0
341	3	2	1	10	1	2	1	0	7	7	–	1.4	5.00	1.0	5.0	3.50	4.0	−3.6	1.50	−3.0
342	4	2	–	5	1	–	1	7	7	0	1	5.0	5.00	3.0	–	–	–	–	–	–
343	4	1	–	8	1	–	2	0	7	-	1	4.2	4.00	3.0	5.0	3.00	–	−0.8	1.00	–
344	4	2	2	8	1	2	1	2	7	7	3	3.8	4.25	3.0	4.2	3.75	5.0	−0.4	0.50	−2.0
345	4	2	2	8	1	2	3	2	7	7	2	3.8	5.00	1.0	5.0	1.00	4.0	−1.4	4.00	−3.0
346	4	1	1	10	2	1	3	2	7	4	1	4.0	4.50	3.5	3.8	2.50	3.0	0.2	2.00	0.5
347	4	2	2	10	1	0	1	1	1	7	1	4.0	5.00	1.0	4.4	3.00	3.0	−0.4	2.00	−2.0
348	2	2	2	9	1	–	1	6	7	5	–	3.8	4.75	2.5	4.6	4.00	5.0	−0.8	0.75	−2.5
349	2	2	2	6	1	2	1	-	–	5	1	–	4.67	-	3.6	2.50	4.0	-	2.17	–
350	3	2	2	8	1	2	1	4	2	4	1	3.8	4.75	1.0	4.2	4.25	4.5	−0.4	0.50	−3.5

No	V2	V3	V4	Know	V7	V8	V9	V10a	V10b	V10c	V11	Health1	Conven1	Financ1	Health2	Conven2	Financ2	Diff1	Diff2	Diff3
351	4	2	2	8	1	–	3	2	5	3	2	5.0	4.75	3.0	5.0	2.50	3.5	0.0	2.25	–0.5
352	4	2	2	6	1	–	3	0	4	7	1	4.2	4.75	2.5	4.8	3.25	4.5	–0.6	1.50	–2.0
353	3	2	2	9	1	1	2	0	7	7	1	3.2	4.50	1.5	4.8	3.50	4.5	–1.6	1.00	–3.0
354	4	2	2	9	1	0	3	0	5	3	2	3.8	4.25	2.0	4.2	3.50	4.5	–0.4	0.75	–2.5
355	4	2	2	10	1	3	1	–	–	–	–	2.2	2.75	3.5	1.6	1.75	1.5	0.6	1.00	2.0
356	3	2	2	6	1	1	1	2	7	7	1	3.8	4.75	2.0	4.8	2.75	4.0	–1.0	2.00	–2.0
357	4	2	1	10	1	2	2	3	6	2	1	4.8	5.00	4.5	2.6	1.00	3.0	2.2	4.00	1.5
358	3	2	2	8	1	2	1	2	7	3	2	4.2	5.00	1.5	3.4	2.50	4.0	0.8	2.50	–2.5
359	4	2	2	8	1	–	3	1	4	3	2	5.0	5.00	4.0	5.0	2.25	2.5	0.0	2.75	1.5
360	4	2	2	6	1	2	2	1	7	7	3	3.0	4.25	3.0	5.0	1.50	5.0	–2.0	2.75	–2.0
361	4	2	2	3	1	0	3	0	5	2	2	3.0	4.25	1.0	3.8	2.50	3.5	–0.8	1.75	–2.5
362	1	2	1	7	1	–	2	0	4	2	1	2.4	5.00	3.0	2.4	5.00	3.0	0.0	0.00	0.0
363	4	2	2	8	1	–	3	0	3	3	2	5.0	5.00	5.0	5.0	2.75	3.5	0.0	2.25	1.5
364	4	2	2	8	1	4	3	3	6	3	3	5.0	5.00	4.5	5.0	2.50	3.5	0.0	2.50	1.0
365	4	2	2	8	1	–	3	0	7	2	2	5.0	5.00	3.5	5.0	2.00	3.5	0.0	3.00	0.0
366	3	2	2	10	2	–	3	5	0	5	–	1.6	2.25	5.0	2.0	3.50	3.0	–0.4	–1.25	2.0
367	4	2	2	8	1	2	3	0	6	1	2	5.0	5.00	5.0	5.0	2.50	3.5	0.0	2.50	1.5
368	2	2	2	10	1	1	1	3	7	7	1	3.2	4.25	3.0	4.8	4.00	4.5	–1.6	0.25	–1.5
369	4	2	2	9	1	1	3	0	5	2	2	3.2	4.00	3.0	3.2	3.00	3.5	0.0	1.00	–0.5
370	3	2	2	2	1	2	3	0	7	4	3	3.6	4.75	1.0	4.8	3.00	3.0	–1.2	1.75	–2.0
371	4	1	1	9	1	2	–	2	5	6	2	4.0	4.25	4.5	4.6	3.75	4.0	–0.6	0.50	0.5
372	2	1	1	8	1	–	3	0	5	4	2	2.6	4.25	3.0	5.0	3.50	3.5	–2.4	0.75	–0.5
373	4	2	2	6	1	–	2	1	5	3	1	3.6	4.50	3.5	4.2	3.00	3.5	–0.6	1.50	0.0
374	4	2	2	8	1	–	3	1	5	3	2	3.6	4.00	3.0	3.8	2.00	4.5	–0.2	2.00	–1.5
375	4	2	1	8	1	2	1	7	0	7	2	4.6	4.50	1.0	4.4	3.50	4.5	0.2	1.00	–3.5
376	2	2	1	6	1	2	1	0	7	7	3	1.2	5.00	1.0	5.0	3.50	3.0	–3.8	1.50	–2.0
377	4	2	1	6	1	3	3	0	7	3	1	2.6	5.00	1.0	5.0	3.00	4.0	–2.4	2.00	–3.0
378	4	1	–	9	2	–	1	0	7	4	1	4.4	4.75	4.0	4.2	2.25	4.0	0.2	2.50	0.0
379	2	2	2	10	1	6	1	7	7	7	3	3.6	4.50	3.0	4.6	1.25	4.0	–1.0	3.25	–1.0
380	3	2	2	3	1	–	3	0	6	4	1	2.2	3.75	2.0	4.6	3.00	3.5	–2.4	0.75	–1.5
381	3	2	2	9	1	2	1	2	6	6	1	2.8	4.50	1.0	4.2	2.75	4.0	–1.4	1.75	–3.0
382	3	2	2	8	1	–	1	3	3	6	2	4.4	4.50	3.0	3.8	3.00	4.0	0.6	1.50	–1.0
383	4	1	–	7	1	2	1	0	7	–	–	4.8	5.00	4.0	5.0	1.00	1.0	–0.2	4.00	3.0
384	4	1	–	8	1	0	2	0	7	5	3	4.6	5.00	4.5	4.2	2.75	3.5	0.4	2.25	1.0
385	3	2	2	10	2	3	3	0	6	5	1	4.4	4.75	4.0	4.2	2.75	3.5	0.2	2.00	0.5
386	4	2	2	9	1	2	1	1	7	6	3	3.8	5.00	1.0	3.8	1.50	4.0	0.0	3.50	–3.0
387	4	1	2	9	1	1	–	–	–	–	1	5.0	4.75	4.5	–	5.00	4.0	–	–0.25	0.5
388	4	2	2	10	1	2	3	0	7	7	1	5.0	4.75	4.5	2.6	1.75	3.0	2.4	3.00	1.5
389	4	2	2	10	1	3	1	0	7	5	–	4.8	5.00	2.5	4.6	1.75	5.0	0.2	3.25	–2.5
390	4	2	2	8	1	0	1	3	7	7	1	4.2	4.75	3.0	4.6	3.50	4.0	–0.4	1.25	–1.0
391	3	2	2	8	1	2	1	1	4	4	2	4.2	5.00	3.0	3.4	3.00	3.0	0.8	2.00	0.0
392	3	2	1	8	1	2	1	0	7	7	2	3.6	4.50	2.0	5.0	3.50	4.0	–1.4	1.00	–2.0
393	4	2	2	8	1	0	1	2	7	7	2	4.6	4.75	3.0	4.40	2.25	4.0	0.20	2.50	–1.0
394	4	2	2	3	2	0	3	1	7	3	2	3.6	3.75	4.0	4.00	3.00	5.0	–0.40	0.75	–1.0
395	4	2	2	7	1	2	1	5	5	6	2	3.4	4.75	1.0	4.80	2.50	4.5	–1.40	2.25	–3.5
396	4	2	1	10	1	0	1	0	7	7	2	4.8	5.00	2.0	5.00	4.50	4.0	–0.20	0.50	–2.0
397	4	2	2	–	1	3	1	5	–	2	2	–	5.00	–	5.00	–	–	–	–	–
398	4	2	2	8	1	1	1	0	7	6	3	4.2	5.00	2.5	4.40	2.25	4.0	–0.20	2.75	–1.5
399	4	2	2	6	1	–	1	2	7	3	2	4.5	5.00	–	1.00	4.00	2.0	3.50	1.00	–
400	4	2	1	7	1	–	1	5	6	6	2	3.5	5.00	–	5.00	3.25	3.5	–1.50	1.75	–

No	V2	V3	V4	Know	V7	V8	V9	V10a	V10b	V10c	V11	Health1	Conven1	Financ1	Health2	Conven2	Financ2	Diff1	Diff2	Diff3
401	4	2	2	9	1	2	1	3	7	7	3	3.6	5.00	3.0	5.00	3.50	4.0	−1.40	1.50	−1.0
402	3	2	–	5	2	2	3	1	5	4	1	4.4	5.00	2.5	5.00	3.00	5.0	−0.60	2.00	−2.5
403	3	2	2	10	1	–	1	4	4	5	1	4.4	4.25	1.5	4.60	3.50	4.0	−0.20	0.75	−2.5
404	3	2	2	10	2	–	3	2	7	7	2	3.8	5.00	2.0	3.60	2.25	4.5	0.20	2.75	−2.5
405	2	2	–	9	1	–	3	1	6	6	2	3.6	5.00	3.0	4.80	3.75	4.0	−1.20	1.25	−1.0
406	4	1	2	7	1	1	1	0	6	5	1	3.4	4.50	2.0	4.80	3.75	4.0	−1.40	0.75	−2.0
407	2	2	2	10	2	–	1	2	4	7	1	4.8	5.00	2.5	4.80	3.75	4.5	0.00	1.25	−2.0
408	3	1	1	2	2	0	1	1	4	4	2	3.6	4.50	3.5	4.20	2.50	4.0	−0.60	2.00	−0.5
409	4	2	2	7	1	–	3	6	7	4	2	3.6	5.00	3.5	3.80	1.00	5.0	−0.20	4.00	−1.5
410	4	2	2	10	1	0	3	0	5	2	1	4.4	5.00	4.5	4.60	2.25	4.5	−0.20	2.75	0.0
411	4	2	2	9	2	0	1	0	7	3	2	4.8	5.00	4.5	4.40	3.25	4.0	0.40	1.75	0.5
412	4	2	1	10	1	–	1	7	7	5	3	4.6	4.75	3.0	4.20	3.25	3.0	0.40	1.50	0.0
413	4	2	2	9	1	2	2	0	3	5	1	4.6	4.75	1.0	4.60	3.00	5.0	0.00	1.75	−4.0
414	3	2	1	4	1	3	2	2	7	7	2	3.6	4.25	1.5	4.20	3.00	4.5	−0.60	1.25	−3.0
415	2	2	1	9	1	2	3	0	7	6	2	2.0	5.00	2.5	5.00	4.25	4.0	−3.00	0.75	−1.5
416	1	2	1	2	1	2	1	0	7	2	1	–	–	–	4.00	2.33	3.0	–	–	–
417	4	2	2	9	2	1	3	0	7	7	1	5.0	5.00	5.0	5.00	5.00	5.0	0.00	0.00	0.0
418	4	1	–	5	2	2	3	1	7	3	1	4.8	4.75	3.0	4.60	3.50	5.0	0.20	1.25	−2.0
419	4	2	2	4	1	2	2	1	4	3	3	3.4	5.00	2.5	3.40	2.00	4.5	0.00	3.00	−2.0
420	3	2	1	9	1	6	2	0	6	5	3	3.2	4.75	1.0	5.00	2.75	4.0	−1.80	2.00	−3.0
421	4	1	–	8	1	3	1	2	5	7	2	4.8	5.00	4.0	3.20	1.50	3.0	1.60	3.50	1.0
422	3	1	2	10	1	–	1	0	2	2	1	4.6	5.00	2.0	2.80	2.25	2.5	1.80	2.75	−0.5
423	4	2	1	7	1	1	1	0	7	7	2	3.2	4.00	3.5	4.00	3.50	3.5	−0.80	0.50	0.0
424	4	1	–	8	1	–	1	3	7	7	1	4.6	5.00	5.0	5.00	3.25	4.0	−0.40	1.75	1.0
425	4	2	2	8	1	–	1	3	6	6	1	4.0	4.25	2.5	3.60	3.25	4.0	0.40	1.00	−1.5
426	4	2	2	8	1	1	1	7	7	7	1	4.0	4.25	5.0	3.40	4.00	1.0	0.60	0.25	4.0
427	4	1	2	1	1	–	2	2	7	7	1	4.0	3.00	3.0	3.20	2.75	3.0	0.80	0.25	0.0
428	3	1	–	10	1	1	3	5	5	3	1	5.0	5.00	5.0	4.20	1.00	4.0	0.80	4.00	1.0
429	4	2	2	8	1	1	1	2	5	1	2	5.0	5.00	3.5	5.00	1.75	4.0	0.00	3.25	−0.5
430	4	1	2	8	1	1	1	1	7	3	2	5.0	5.00	3.0	4.20	1.75	3.0	0.80	3.25	0.0
431	3	2	2	9	1	1	1	0	7	7	2	3.4	2.25	4.5	2.40	3.00	2.5	1.00	−0.75	2.0
432	4	2	2	9	1	1	3	0	7	4	1	2.6	2.00	4.5	3.60	4.00	1.0	−1.00	−2.00	3.5
433	4	2	2	7	1	-	1	0	7	7	2	2.4	2.75	4.0	2.20	3.00	2.0	0.20	−0.25	2.0
434	4	2	2	8	1	-	1	0	7	7	2	3.0	3.50	4.0	3.20	3.00	3.0	−0.20	0.50	1.0
435	4	2	2	6	1	1	1	0	7	7	2	3.8	3.00	4.0	4.20	3.25	2.5	−0.40	−0.25	1.5
436	4	2	2	8	1	1	3	1	1	2	–	4.2	4.00	5.0	3.20	3.75	2.0	1.00	0.25	3.0
437	1	2	1	3	1	0	3	0	3	7	1	4.0	2.00	5.0	3.40	2.00	3.0	0.60	0.00	2.0
438	3	2	2	9	2	-	1	0	7	2	1	3.6	2.50	2.5	-	-	-	-	-	-
439	1	2	1	6	1	1	1	0	7	7	1	–	–	–	3.20	2.75	2.0	–	–	–
440	4	1	–	8	1	4	3	0	7	7	1	3.4	2.75	2.5	5.00	5.00	5.0	-1.60	−2.25	−2.5
441	1	2	2	7	1	–	1	7	7	5	1	–	–	–	2.40	2.25	3.0	–	–	–
442	4	2	2	8	1	2	1	2	7	6	2	4.2	4.75	1.5	4.20	3.00	3.5	0.00	1.75	−2.0
443	2	2	1	10	1	5	2	0	7	7	1	2.8	3.00	4.5	5.00	5.00	–	−2.20	−2.00	–
444	4	2	2	5	2	-	3	0	7	7	1	4.0	4.00	4.5	5.00	5.00	5.0	−1.00	-1.00	−0.5
445	4	2	1	8	1	2	2	0	7	4	1	3.8	4.50	1.0	4.80	4.25	4.0	−1.00	0.25	−3.0
446	3	1	1	8	2	1	3	2	5	5	2	2.8	5.00	3.0	1.20	3.75	3.5	1.60	1.25	−0.5
447	4	1	2	9	1	2	1	7	7	7	-	4.2	4.50	4.0	3.00	1.00	3.0	1.20	3.50	1.0
448	4	1	2	2	1	4	1	4	7	-	2	4.2	5.00	3.0	3.67	3.00	3.0	0.53	2.00	0.0
449	4	1	1	2	2	–	1	0	6	7	3	–	–	–	5.0	5.00	–	–	–	–
450	4	2	2	8	1	0	3	0	7	4	1	5.0	5.00	3.5	4.0	3.50	4.0	1.0	1.50	−0.5

No	V2	V3	V4	Know	V7	V8	V9	V10a	V10b	V10c	V11	Health1	Conven1	Financ1	Health2	Conven2	Financ2	Diff1	Diff2	Diff3
451	4	1	2	7	2	0	3	1	5	5	2	5.0	5.00	5.0	4.8	4.00	4.5	0.2	1.00	0.5
452	4	2	2	7	1	2	1	2	6	7	2	4.8	4.75	4.0	4.2	2.25	4.0	0.6	2.50	0.0
453	4	2	1	4	1	3	1	0	3	3	2	2.8	3.75	2.5	4.0	3.00	4.0	−1.2	0.75	−1.5
454	4	2	1	8	1	–	–	3	3	0	–	4.2	3.75	5.0	3.2	3.50	4.0	1.0	0.25	1.0
455	4	2	1	10	1	–	1	0	7	7	3	3.6	4.75	2.5	3.6	4.00	3.5	0.0	0.75	−1.0
456	3	2	2	8	1	–	–	0	7	7	3	4.0	4.50	1.0	4.6	2.50	3.0	−0.6	2.00	−2.0
457	3	2	1	6	1	–	3	7	7	7	3	4.8	5.00	3.0	4.8	4.00	5.0	0.0	1.00	−2.0
458	2	2	1	8	1	2	1	3	7	7	1	3.2	4.50	2.0	4.4	3.25	4.0	−1.2	1.25	−2.0
459	2	2	1	9	–	–	–	–	–	–	–	3.6	3.75	2.0	4.4	4.00	5.0	−0.8	−0.25	−3.0
460	3	2	2	9	2	–	1	0	7	7	2	5.0	5.00	5.0	4.8	3.00	4.0	0.2	2.00	1.0
461	4	2	2	6	1	2	2	–	5	4	2	3.8	4.75	4.0	5.0	3.50	4.5	−1.2	1.25	−0.5
462	4	–	–	8	1	1	2	2	6	1	2	4.8	5.00	4.0	3.8	2.00	3.0	1.0	3.00	1.0
463	4	2	1	9	1	0	1	4	4	7	1	3.4	5.00	5.0	3.0	3.00	5.0	0.4	2.00	0.0
464	3	2	2	7	1	0	1	1	7	7	2	4.6	4.75	3.5	3.6	3.00	2.5	1.0	1.75	1.0
465	4	2	2	8	1	–	1	–	7	7	2	4.0	4.50	4.0	3.8	3.25	3.5	0.2	1.25	0.5
466	3	2	2	10	1	2	1	7	7	7	1	4.0	4.50	2.5	4.6	3.75	5.0	−0.6	0.75	−2.5
467	2	2	2	5	1	2	2	1	7	3	2	3.2	3.50	3.0	4.6	3.75	5.0	−1.4	−0.25	−2.0
468	2	2	1	7	1	6	2	1	7	7	2	2.6	4.25	1.0	5.0	3.25	4.0	−2.4	1.00	−3.0
469	4	2	1	6	1	1	2	0	6	–	2	4.4	4.75	1.0	3.6	2.25	5.0	0.8	2.50	−4.0
470	4	2	2	10	1	1	1	3	7	7	1	4.6	4.75	2.0	4.4	1.25	4.5	0.2	3.50	−2.5
471	4	2	2	3	1	–	1	1	7	6	2	3.4	4.50	2.0	4.8	2.50	5.0	−1.4	2.00	−3.0
472	3	2	1	8	1	3	2	1	6	6	3	4.6	4.00	1.0	4.8	3.00	4.5	−0.2	1.00	−3.5
473	4	2	2	3	1	2	2	7	7	7	1	4.0	4.75	1.0	4.6	2.00	4.0	−0.6	2.75	−3.0
474	4	2	1	8	1	0	3	0	7	7	3	4.2	5.00	3.0	5.0	4.25	3.0	−0.8	0.75	0.0
475	4	1	–	8	1	2	3	0	7	2	1	4.8	5.00	5.0	4.4	2.00	3.5	0.4	3.00	1.5
476	4	2	2	5	1	0	3	2	7	3	2	3.6	5.00	3.0	4.4	2.25	4.0	−0.8	2.75	−1.0
477	4	1	–	8	2	–	–	–	–	–	–	–	5.00	–	5.0	–	–	–	–	–
478	4	1	1	7	1	–	1	0	7	5	2	–	5.00	–	5.0	5.00	–	–	0.00	–
479	4	2	2	7	2	–	–	7	7	7	1	1.2	1.50	2.0	1.8	2.25	2.0	−0.6	−0.75	0.0
480	4	1	1	8	1	2	3	0	7	7	1	4.8	5.00	2.5	4.0	3.75	4.5	0.8	1.25	−2.0
481	3	2	1	10	1	–	1	0	7	7	2	4.4	5.00	1.0	4.6	4.75	4.0	−0.2	0.25	−3.0
482	3	2	1	4	1	2	1	2	7	6	2	3.4	4.00	2.0	4.2	4.00	4.0	−0.8	0.00	−2.0
483	2	2	1	6	1	2	3	1	6	7	2	2.6	4.50	1.0	4.8	3.75	4.5	−2.2	0.75	−3.5
484	1	2	1	10	1	–	1	0	7	7	–	–	–	–	4.4	5.00	4.5	–	–	–
485	3	1	1	8	1	2	2	1	6	6	3	2.4	4.00	3.0	4.8	3.50	5.0	−2.4	0.50	−2.0
486	3	2	1	7	1	2	2	–	–	–	2	3.4	4.50	2.5	4.0	3.50	4.0	−0.6	1.00	−1.5
487	4	1	2	9	1	0	1	0	7	7	1	4.4	4.75	3.5	3.0	1.50	2.5	1.4	3.25	1.0
488	4	2	2	6	2	–	3	3	5	–	2	4.6	4.50	3.5	3.4	2.00	4.5	1.2	2.50	−1.0
489	4	2	2	8	2	1	1	7	7	7	1	–	5.00	4.0	5.0	4.00	5.0	–	1.00	−1.0
490	4	2	2	8	1	–	3	0	7	7	1	3.4	5.00	1.0	4.8	4.50	4.5	−1.4	0.50	−3.5
491	2	1	2	10	1	3	2	0	7	7	1	2.0	3.25	1.0	3.8	3.00	2.0	−1.8	0.25	−1.0
492	4	2	2	8	2	–	3	2	5	5	2	4.8	4.75	4.0	3.2	2.25	4.0	1.6	2.50	0.0
493	4	2	2	9	1	3	3	3	7	6	2	3.2	3.75	1.5	4.6	3.50	4.5	−1.4	0.25	−3.0
494	4	2	2	6	1	–	1	0	7	7	1	4.4	4.50	3.5	4.8	3.25	4.0	−0.4	1.25	−0.5
495	4	2	2	5	1	0	3	0	5	2	2	4.6	5.00	3.0	5.0	3.00	4.0	−0.4	2.00	−1.0
496	2	2	1	9	1	2	1	7	2	7	1	3.0	4.50	1.0	4.8	3.33	4.0	−1.8	1.17	−3.0
497	3	2	2	7	1	–	1	3	6	6	2	3.8	4.75	2.5	5.0	4.00	4.5	−1.2	0.75	−2.0
498	4	1	2	10	1	–	1	7	0	5	1	4.2	5.00	4.0	2.8	2.50	3.0	1.4	2.50	1.0
499	2	1	–	6	1	–	1	7	0	7	1	1.6	2.00	3.0	5.0	4.00	2.0	−3.4	−2.00	1.0
500	2	2	1	10	1	–	1	7	0	7	1	4.4	5.00	3.5	4.4	4.50	4.5	0.0	0.50	−1.0

No	V2	V3	V4	Know	V7	V8	V9	V10a	V10b	V10c	V11	Health1	Conven1	Financ1	Health2	Conven2	Financ2	Diff1	Diff2	Diff3
501	4	2	–	7	1	2	1	0	7	5	2	4.0	4.50	3.5	2.2	3.00	4.0	1.8	1.50	–0.5
502	4	2	2	5	1	3	3	7	7	3	1	3.0	5.00	4.0	4.6	2.75	3.0	–1.6	2.25	1.0
503	4	2	2	10	1	2	3	2	7	5	2	4.0	4.75	3.0	4.4	2.75	4.0	–0.4	2.00	–1.0
504	4	2	2	2	1	2	1	0	7	4	2	3.4	4.50	3.0	3.6	4.00	2.5	–0.2	0.50	0.5
505	4	2	2	9	1	2	3	0	6	2	3	4.0	4.75	2.5	4.4	3.00	4.5	–0.4	1.75	–2.0
506	4	2	1	9	1	2	3	0	6	6	1	4.0	4.75	4.0	5.0	2.50	5.0	–1.0	2.25	–1.0
507	4	1	–	2	1	2	1	0	7	5	3	3.4	4.50	5.0	3.8	1.75	3.0	–0.4	2.75	2.0
508	4	2	1	8	1	2	1	0	7	7	3	4.4	4.50	3.0	3.8	2.00	5.0	0.6	2.50	–2.0
509	3	2	1	10	2	–	3	0	4	3	2	3.0	4.50	2.0	5.0	3.00	3.5	–2.0	1.50	–1.5
510	4	2	2	9	1	2	–	1	4	4	3	4.0	3.75	4.0	4.0	4.25	3.0	0.0	–0.50	1.0
511	2	2	1	10	1	2	2	0	7	7	3	4.0	5.00	1.5	5.0	5.00	1.0	–1.0	0.00	0.5
512	4	1	2	2	1	0	3	0	7	4	3	4.2	4.75	4.0	4.0	2.50	3.5	0.2	2.25	0.5
513	4	1	1	7	2	–	3	0	4	0	–	4.4	5.00	4.0	2.2	2.00	1.0	2.2	3.00	3.0
514	4	2	2	8	1	–	1	0	7	7	3	1.6	3.00	–	1.4	3.00	3.0	0.2	0.00	–
515	4	2	2	9	1	2	3	0	7	5	2	–	5.00	1.0	1.0	5.00	1.0	–	0.00	0.0
516	4	1	–	10	1	2	1	0	7	7	–	3.0	5.00	3.5	4.0	3.25	4.0	–1.0	1.75	–0.5
517	4	2	2	9	1	2	1	0	7	4	2	3.8	5.00	3.0	4.4	2.25	4.5	–0.6	2.75	–1.5
518	3	2	2	8	1	3	3	1	7	7	3	3.6	4.50	1.5	4.2	2.50	4.0	–0.6	2.00	–2.5
519	4	2	–	8	1	2	1	0	7	7	3	4.4	5.00	2.5	3.8	2.75	3.0	0.6	2.25	–0.5
520	4	2	2	9	1	–	3	0	2	1	2	3.8	5.00	2.0	4.2	2.25	5.0	–0.4	2.75	–3.0
521	–	2	2	8	1	1	2	3	7	7	3	–	–	–	4.2	3.00	3.5	–	–	–
522	3	2	2	9	1	1	3	3	7	7	1	4.4	5.00	1.0	4.6	3.25	5.0	–0.2	1.75	–4.0
523	4	1	2	8	1	–	1	1	6	7	2	4.0	4.75	2.5	3.4	2.75	4.0	0.6	2.00	–1.5
524	3	1	1	8	2	–	1	7	4	7	1	4.2	4.50	2.0	4.6	4.00	4.0	–0.4	0.50	–2.0
525	4	2	2	–	1	3	1	0	7	5	3	–	5.00	–	3.0	1.00	–	–	4.00	–
526	1	–	–	2	2	–	1	6	5	7	2	3.0	3.00	3.0	4.0	3.75	3.0	–1.0	–0.75	0.0
527	4	2	2	9	2	–	3	1	7	7	2	4.6	5.00	3.0	4.6	3.50	2.5	0.0	1.50	0.5
528	3	2	1	9	1	–	1	0	5	7	2	3.8	5.00	1.0	5.0	3.50	5.0	–1.2	1.50	–4.0
529	4	2	2	7	1	2	3	4	7	2	2	3.0	4.00	1.5	3.2	2.50	4.0	–0.2	1.50	–2.5
530	4	2	1	4	1	5	1	7	7	4	3	2.6	4.50	3.0	3.8	2.50	3.5	–1.2	2.00	–0.5
531	4	2	1	6	1	3	1	0	7	7	2	3.6	4.25	1.0	4.8	3.75	4.0	–1.2	0.50	–3.0
532	4	2	2	10	1	2	1	5	0	7	3	5.0	5.00	4.0	4.6	3.25	4.0	0.4	1.75	0.0
533	4	1	–	8	1	3	2	1	6	7	–	5.0	5.00	4.0	2.8	1.25	5.0	2.2	3.75	–1.0
534	4	2	2	10	1	2	1	0	7	7	2	5.0	5.00	4.5	5.0	3.25	5.0	0.0	1.75	–0.5
535	4	2	2	8	2	1	3	0	7	7	1	5.0	4.00	3.0	5.0	2.00	5.0	0.0	2.00	–2.0
536	4	2	2	10	1	–	2	0	7	7	3	3.8	3.50	2.0	3.2	2.75	3.0	0.6	0.75	–1.0
537	4	1	–	10	1	3	1	0	7	7	2	5.0	5.00	4.0	4.6	1.50	3.0	0.4	3.50	1.0
538	4	2	2	8	1	4	3	2	4	3	2	5.0	5.00	4.0	1.0	–	–	4.0	–	–
539	2	2	1	10	1	–	3	1	6	7	3	3.6	5.00	3.5	4.0	3.50	4.5	–0.4	1.50	–1.0
540	4	2	2	9	–	–	3	3	7	4	1	4.4	5.00	4.0	4.6	2.25	4.0	–0.2	2.75	0.0
541	4	2	2	9	1	–	3	3	7	4	1	4.4	5.00	4.0	4.6	2.25	4.0	–0.2	2.75	0.0
542	2	2	1	9	1	4	2	2	6	5	3	3.0	4.75	2.0	4.4	3.25	4.0	–1.4	1.50	–2.0
543	2	2	2	10	1	3	1	0	7	7	2	4.2	3.75	1.0	5.0	4.33	4.0	–0.8	–0.58	–3.0

Appendix A

Data Set 3

This data set contains information on 440 babies and their mothers. The data was obtained from a gynaecologist. The following variables are given in the data set:

No	=	The observation number.	
ID	=	A number by which the mother can be identified on file. The first letter corresponds to the first letter of the mother's surname.	
Mage	=	The mother's age (in years).	
Gr	=	Gravida, the number of pregnancies the mother has had.	
Bage	=	The birth age of the baby in weeks. Normally the birth age is 40 weeks.	
Gender	=	The baby's gender	1 = Boy 2 = Girl.
Bmass	=	The baby's birth mass in kilograms.	
Delmeth	=	The method of delivery	1 = Normal birth 2 = Caesarian section.
Onset	=	The onset of labour	1 = Spontaneous 2 = Induction.

No	ID	Mage	Gr	Bage	Gender	Bmass	Delmeth	Onset
1	A001P	31	1	39.0	2	3.20	1	2
2	A002P	31	3	42.7	1	3.70	2	1
3	A001T	32	5	38.9	2	4.24	2	2
4	A010T	29	2	40.3	2	3.30	1	1
5	A019T	28	2	39.1	1	3.16	1	1
6	A020T	29	3	37.9	1	3.38	2	2
7	B008P	23	1	39.0	2	3.30	1	2
8	B009P	28	2	37.7	1	2.90	2	1
9	B014P	21	3	41.4	2	3.19	1	1
10	B019P	23	1	40.9	1	2.50	1	1
11	B004T	25	3	36.4	2	3.95	2	2
12	B005T	31	5	37.7	1	3.50	1	2
13	B007T	22	2	40.7	2	2.76	1	1
14	B008T	21	2	41.3	1	3.74	1	1
15	B011T	29	1	39.6	2	3.34	2	2
16	B013T	24	2	37.9	2	3.20	1	2
17	B016T	33	3	38.3	1	3.37	2	1
18	B018T	27	1	39.9	2	3.73	1	1
19	B019T	25	2	39.3	1	3.70	1	2
20	B020T	20	2	38.7	2	2.78	1	1
21	B022T	22	1	40.0	2	3.23	1	1
22	B025T	25	5	41.7	2	3.37	1	1
23	B030T	21	2	40.0	1	3.34	1	1
24	B031T	22	1	40.3	2	3.22	2	2
25	B032T	23	1	39.1	1	3.50	2	1
26	B033T	26	2	40.1	2	3.30	2	1
27	B036T	25	2	40.7	1	3.34	2	1
28	B041T	21	2	35.4	1	2.82	1	1
29	B052T	25	4	37.7	1	3.57	1	2
30	B060T	27	2	38.3	1	3.34	1	1
31	B069T	27	1	39.0	2	3.60	1	2
32	B070T	20	2	38.1	1	3.50	1	2
33	B073T	31	1	38.7	1	2.96	1	1
34	B075T	40	6	37.7	1	3.66	1	2
35	B085T	22	2	41.4	2	2.90	1	1
36	B089T	27	1	37.0	1	3.18	2	1
37	B090T	23	4	39.6	1	3.30	1	1
38	B095T	19	1	39.6	1	3.40	1	1
39	B096T	20	1	39.0	2	3.50	1	2
40	B098T	22	1	39.7	2	3.10	2	2
41	B100T	25	2	38.9	2	3.22	1	1
42	B103T	22	1	42.0	2	3.70	1	1
43	B104T	32	3	39.1	2	3.04	1	1
44	B107T	25	4	37.9	1	3.52	1	2
45	B110T	21	1	41.3	2	4.30	2	2
46	B112T	23	1	40.3	1	4.10	1	1
47	B115T	32	6	39.6	1	4.30	2	1
48	B117T	22	3	39.9	2	3.44	1	2
49	B119T	39	2	36.3	1	3.40	2	2
50	B128T	18	2	39.4	1	3.70	1	1
51	B139T	22	3	40.1	2	2.62	2	1
52	B141T	18	1	41.4	2	3.52	1	1
53	B146T	27	2	37.1	1	3.13	1	1
54	C010P	25	1	38.0	1	3.36	1	2
55	C002T	25	2	41.0	1	3.00	1	1
56	C003T	26	2	36.4	1	3.50	1	1
57	C005T	27	3	35.3	2	2.90	2	1
58	C010T	20	3	37.6	2	3.30	1	2
59	C017T	27	5	38.4	1	3.30	1	2
60	C018T	20	1	35.4	2	3.32	1	1
61	C021T	32	3	40.1	2	3.20	1	1
62	C025T	18	1	38.3	1	3.10	1	1
63	C027T	30	4	37.4	2	2.90	1	2
64	C031T	22	2	40.6	2	3.50	2	2
65	C034T	22	2	38.0	1	3.10	1	1
66	C037T	24	2	39.4	2	3.25	1	1
67	C042T	26	2	36.3	1	2.80	1	2
68	C043T	18	1	38.6	1	3.30	1	1
69	C049T	24	5	39.3	1	3.52	1	2
70	C052T	24	4	39.4	1	3.50	1	2
71	C053T	27	2	39.3	1	3.60	1	1
72	C059T	25	5	38.9	2	3.10	1	1
73	D002T	26	2	35.9	2	2.78	1	2
74	D003T	32	4	38.3	1	3.80	1	1
75	D005T	30	3	39.4	2	3.15	1	1
76	D009T	25	4	39.6	1	3.85	1	1
77	D014T	34	3	38.6	1	3.60	1	1
78	D022T	29	3	39.1	1	4.20	1	1
79	D035T	24	3	36.7	1	2.50	1	1
80	D046T	20	2	39.3	2	3.30	1	1
81	D049T	29	3	40.0	2	3.36	1	1
82	D050T	20	1	38.4	1	3.14	1	1
83	D054T	23	4	40.4	2	4.20	1	2
84	D060T	20	1	37.4	2	3.00	1	1
85	D062T	31	3	37.7	2	3.15	1	1
86	D065T	23	1	39.9	2	3.50	1	1
87	D076T	25	1	37.0	1	3.30	1	1
88	D079T	22	1	39.0	1	3.70	1	1
89	D082T	29	2	40.3	1	4.50	1	1
90	D083T	32	6	40.4	2	3.20	1	1
91	D088T	22	2	39.1	2	4.40	1	1

No	ID	Mage	Gr	Bage	Gender	Bmass	Delmeth	Onset
92	D091T	17	1	40.4	2	3.40	2	1
93	D096T	35	7	38.9	1	2.80	1	1
94	D097T	21	2	36.9	2	3.60	1	2
95	D103T	21	1	39.6	2	3.60	1	2
96	D106T	24	1	38.9	2	3.18	1	1
97	D109T	16	1	38.4	1	3.40	1	1
98	E001P	25	4	38.0	1	3.59	1	2
99	E002P	26	1	40.6	2	2.88	1	2
100	E004P	32	6	39.0	2	4.30	1	2
101	E006P	31	2	38.0	2	3.60	1	2
102	E008P	26	1	40.3	2	3.40	1	1
103	E005T	20	1	38.4	2	3.70	1	2
104	E006T	24	2	40.4	2	2.62	2	1
105	E008T	21	2	38.7	2	3.50	1	1
106	E017T	21	1	37.7	2	2.40	1	1
107	E018T	40	7	34.7	1	2.50	1	1
108	E021T	23	1	39.7	2	3.40	2	2
109	E022T	23	2	39.7	2	3.34	1	2
110	E023T	23	2	37.7	1	3.60	1	2
111	E029T	19	1	38.3	2	3.35	2	2
112	F004T	25	4	37.6	2	3.50	1	1
113	F013T	24	2	37.9	2	2.60	1	1
114	F024T	23	2	40.0	2	3.26	1	2
115	F026T	29	2	42.7	2	3.80	1	1
116	F029T	24	1	37.6	1	3.74	2	2
117	G007P	26	2	41.0	2	3.20	1	1
118	G001T	27	1	35.1	1	2.44	1	1
119	G004T	30	2	40.0	2	3.79	1	1
120	G006T	18	1	39.4	1	3.40	1	1
121	G011T	24	3	37.3	2	3.50	1	1
122	G013T	27	1	39.6	2	2.88	1	1
123	G020T	23	2	39.9	1	3.34	1	1
124	G025T	23	2	38.7	2	3.20	2	2
125	G027T	20	2	38.6	2	3.05	2	1
126	G032T	27	2	39.3	1	3.84	2	2
127	G033T	24	1	39.6	2	3.30	1	2
128	G036T	31	3	39.1	1	3.40	1	2
129	G038T	24	2	38.6	1	3.10	1	1
130	G040T	38	4	39.4	2	3.25	1	2
131	G042T	25	1	39.6	1	3.06	1	2
132	G045T	19	1	40.4	1	4.00	1	1

No	ID	Mage	Gr	Bage	Gender	Bmass	Delmeth	Onset
133	H002P	29	2	39.6	1	3.54	2	2
134	H005P	22	1	40.6	1	4.20	1	2
135	H008P	25	2	40.0	2	3.60	1	1
136	H010P	29	1	39.9	1	3.09	1	1
137	H001T	26	2	38.1	2	3.00	1	1
138	H002T	23	2	37.4	1	3.82	2	1
139	H005T	26	1	38.9	2	3.20	1	1
140	H011T	21	1	42.4	1	2.95	1	1
141	H020T	25	3	38.0	1	4.28	1	1
142	H021T	22	2	36.9	1	3.45	1	1
143	H027T	21	1	38.1	1	3.20	1	1
144	H028T	24	1	40.0	2	3.20	2	2
145	H032T	20	2	40.6	1	3.58	1	1
146	H035T	30	3	37.9	2	2.44	1	1
147	H036T	24	4	37.1	2	3.50	1	1
148	H040T	26	2	38.3	2	3.65	1	1
149	H041T	21	3	40.4	2	3.50	1	2
150	H054T	24	2	42.4	1	3.36	1	1
151	H055T	24	1	38.6	1	3.10	1	1
152	H060T	24	1	40.6	1	3.00	1	2
153	H063T	25	1	39.6	1	4.25	2	1
154	H064T	30	2	37.9	2	2.84	1	1
155	H065T	20	2	40.0	1	3.64	1	1
156	H066T	28	1	38.6	2	3.80	2	2
157	H071T	25	1	40.4	2	3.92	2	1
158	J002P	28	2	38.9	2	3.53	1	1
159	J005P	25	2	37.3	1	3.18	2	2
160	J002T	25	2	40.3	2	4.10	1	1
161	J010T	26	2	38.3	2	2.40	1	1
162	J013T	33	3	38.4	2	3.35	2	1
163	J014T	24	2	41.4	1	2.70	2	1
164	J015T	28	5	39.4	2	3.13	1	1
165	J021T	19	1	40.1	2	3.20	1	1
166	J023T	30	1	36.7	1	3.04	1	1
167	J031T	26	2	39.7	1	3.90	1	1
168	J035T	27	1	39.6	1	3.00	2	1
169	J036T	30	3	42.3	1	3.86	2	2
170	J038T	23	2	38.6	2	3.50	1	1
171	K003P	24	1	37.9	2	3.00	1	1
172	K013P	28	2	37.3	1	2.97	2	1
173	K010T	29	3	38.9	1	3.59	1	1
174	K023T	29	3	38.7	2	3.18	1	2
175	K024T	24	2	38.3	2	3.38	2	1
176	K031T	21	1	40.9	2	3.62	1	2
177	K033T	23	2	38.4	2	3.06	1	1

No	ID	Mage	Gr	Bage	Gender	Bmass	Delmeth	Onset
178	K034T	25	1	39.7	1	3.42	1	1
179	K035T	29	4	40.3	2	3.70	1	1
180	K039T	25	3	38.6	2	3.47	1	2
181	K042T	26	4	40.7	2	3.75	1	2
182	K044T	26	4	38.6	1	3.45	1	2
183	K047T	27	4	36.6	1	3.12	1	1
184	K055T	24	2	36.4	2	2.85	1	1
185	K057T	25	2	40.3	1	3.71	2	1
186	K060T	24	2	38.7	2	3.10	1	2
187	K062T	23	3	38.4	1	3.20	1	1
188	L001P	39	4	36.6	2	3.10	2	2
189	L005P	28	2	38.9	2	3.29	1	1
190	L007P	28	4	37.1	2	3.75	1	1
191	L008P	25	2	38.6	2	3.61	1	2
192	L002T	23	1	40.6	1	3.37	2	1
193	L006T	26	1	38.6	1	3.70	1	1
194	L007T	22	6	42.3	2	3.00	1	1
195	L011T	26	2	37.7	1	3.40	1	1
196	L012T	22	3	42.3	2	3.13	1	1
197	L013T	24	2	42.7	2	3.30	2	1
198	L014T	33	3	39.3	1	3.80	1	1
199	L015T	26	1	41.3	2	3.56	1	1
200	L018T	23	2	38.4	1	4.20	1	1
201	L021T	26	2	41.0	2	3.29	1	1
202	L024T	19	2	40.6	2	2.80	1	2
203	L025T	24	1	40.6	1	2.92	1	1
204	L031T	20	1	37.4	2	3.00	1	1
205	L032T	25	1	40.9	2	3.74	1	1
206	L041T	22	2	36.1	1	2.90	1	1
207	L050T	21	3	42.7	2	2.78	1	1
208	L053T	28	3	36.7	1	2.08	2	1
209	L059T	27	3	40.3	2	3.20	1	1
210	L063T	27	5	40.3	1	3.78	2	2
211	L083T	24	2	41.4	2	3.40	1	1
212	M001P	21	1	35.4	1	2.50	1	1
213	M002P	19	1	37.6	1	3.36	1	1
214	M009P	26	1	37.1	2	3.55	1	1
215	M013P	26	1	37.1	2	3.15	1	1
216	M014P	29	2	39.6	1	3.63	1	1
217	M017P	29	2	39.3	2	3.10	1	2
218	M021P	26	3	40.7	1	3.00	1	1
219	M022P	20	1	40.0	1	3.46	1	2
220	M005T	18	1	36.7	2	3.20	1	1
221	M006T	31	3	40.4	1	3.52	2	1

No	ID	Mage	Gr	Bage	Gender	Bmass	Delmeth	Onset
222	M007T	27	2	37.6	2	3.08	2	1
223	M009T	27	4	40.7	2	3.35	1	1
224	M010T	27	2	39.0	1	3.18	1	1
225	M013T	28	3	38.3	2	3.67	1	1
226	M023T	22	3	41.1	1	3.09	1	1
227	M031T	23	1	40.0	1	3.85	2	2
228	M035T	25	1	40.9	1	3.66	1	1
229	M040T	19	1	38.1	1	3.10	1	1
230	M044T	26	1	38.4	1	3.90	1	1
231	M046T	27	6	37.7	1	3.85	1	2
232	M049T	24	1	40.6	2	4.25	1	2
233	M052T	20	1	39.7	2	2.80	2	1
234	M056T	24	2	39.4	2	3.60	2	1
235	M057T	31	2	37.4	1	3.30	2	2
236	M058T	33	3	38.1	2	2.94	1	1
237	M059T	22	1	37.7	2	2.70	2	1
238	M061T	27	3	39.3	2	3.85	2	1
239	M072T	32	4	38.3	2	3.75	1	1
240	M073T	21	2	39.3	1	2.70	1	2
241	M074T	22	1	41.4	1	3.80	1	1
242	M076T	19	1	39.6	1	3.40	1	1
243	M077T	25	2	36.4	1	2.70	1	1
244	M079T	22	2	37.7	2	2.80	1	1
245	M080T	27	2	37.1	1	2.70	1	1
246	M089T	30	3	37.0	2	2.90	1	2
247	M091T	22	3	38.1	2	2.88	1	1
248	M092T	25	2	38.1	2	3.20	1	1
249	N003P	25	1	39.3	2	3.40	1	2
250	N011P	24	1	39.3	1	3.70	1	1
251	N002T	25	4	40.6	2	3.60	1	1
252	N004T	20	3	37.6	2	2.80	2	1
253	N008T	19	1	37.1	1	2.92	2	1
254	N014T	20	2	39.6	2	3.48	2	2
255	N015T	24	2	36.1	1	2.70	1	1
256	N026T	27	2	38.3	2	3.30	1	2
257	N028T	29	1	41.1	2	3.36	1	1
258	N030T	25	1	39.7	1	3.00	1	1
259	N033T	26	1	40.7	1	3.08	1	1
260	N041T	31	4	37.4	1	3.54	1	2

No	ID	Mage	Gr	Bage	Gender	Bmass	Delmeth	Onset
261	O002P	26	1	40.1	1	3.10	1	2
262	O003T	24	4	36.7	1	3.50	1	1
263	O005T	29	2	40.0	2	2.80	1	2
264	O010T	20	1	38.1	2	3.20	1	1
265	O020T	29	2	38.4	1	3.90	1	2
266	O023T	28	4	38.7	2	3.60	1	1
267	O027T	26	3	38.6	1	3.38	1	1
268	P005P	32	1	36.4	1	3.17	1	1
269	P012P	24	1	39.1	1	3.48	2	2
270	P021P	23	1	38.6	2	2.60	1	1
271	P025P	29	1	39.6	2	3.40	1	1
272	P003T	24	1	38.3	1	3.30	1	1
273	P006T	35	3	39.0	1	3.10	1	1
274	P007T	21	1	38.3	1	4.00	2	1
275	P013T	23	1	38.7	2	3.26	1	1
276	P018T	18	1	37.0	1	3.10	2	1
277	P019T	22	2	38.3	1	3.40	1	1
278	P027T	36	3	37.3	1	4.03	1	2
279	P028T	30	3	39.7	2	3.70	1	1
280	P030T	29	3	37.0	1	3.28	1	1
281	P034T	22	1	40.1	2	3.12	1	1
282	P038T	27	2	37.7	1	2.90	1	2
283	P040T	23	2	38.4	2	3.40	1	2
284	P043T	22	4	42.3	1	3.74	1	2
285	P046T	28	5	40.7	2	3.60	1	1
286	P050T	19	2	38.3	1	3.35	1	1
287	P051T	30	3	40.3	1	4.50	1	2
288	P053T	23	2	39.1	2	3.52	1	1
289	P054T	20	1	39.7	2	3.30	1	1
290	P056T	27	4	37.1	1	3.35	2	2
291	P060T	20	1	43.9	2	2.62	1	1
292	P062T	26	1	39.9	1	4.20	1	2
293	R003P	20	1	39.1	1	4.15	2	1
294	R004T	32	5	42.4	2	3.04	2	1
295	R008T	27	2	39.4	1	3.70	1	1
296	R010T	37	3	36.9	1	3.30	2	1
297	R011T	30	3	40.9	1	4.32	2	2
298	R013T	28	3	38.1	1	4.23	1	2
299	R014T	27	3	41.4	1	4.02	1	2
300	R016T	24	2	34.7	1	2.97	2	1
301	R017T	21	2	39.1	1	3.28	1	2
302	R019T	25	3	40.1	2	3.00	1	1
303	R021T	23	4	35.6	1	2.49	1	2
304	R023T	23	2	39.0	1	2.60	1	1
305	R024T	24	1	43.9	2	3.44	2	1

No	ID	Mage	Gr	Bage	Gender	Bmass	Delmeth	Onset
306	R029T	21	1	39.7	1	3.03	1	1
307	R032T	23	3	39.4	1	3.07	1	1
308	R047T	34	4	37.7	2	3.68	1	1
309	R052T	33	4	39.7	1	4.10	1	2
310	R058T	21	2	38.6	2	2.10	1	1
311	S001P	35	2	35.7	1	3.20	1	2
312	S004P	25	2	36.0	1	3.00	1	1
313	S007P	22	1	38.1	1	3.20	1	2
314	S010P	32	3	38.1	2	3.50	1	2
315	S011P	22	2	39.0	2	3.65	1	1
316	S013P	23	1	38.0	1	3.66	1	2
317	S009T	28	4	38.1	1	2.90	1	2
318	S014T	22	1	39.4	1	3.24	1	1
319	S016T	25	1	39.4	2	3.23	1	1
320	S023T	21	2	39.9	1	3.00	1	1
321	S032T	22	1	40.1	2	3.57	1	2
322	S035T	21	1	41.4	1	3.44	1	1
323	S038T	23	1	38.4	1	2.71	1	1
324	S039T	22	2	37.7	1	2.70	2	1
325	S043T	29	4	38.0	1	4.20	1	1
326	S046T	24	2	40.6	1	3.85	1	1
327	S048T	28	2	38.6	1	2.77	1	1
328	S049T	18	1	41.4	2	2.94	1	1
329	S050T	18	1	39.3	1	3.04	1	1
330	S054T	28	2	37.7	1	4.17	1	1
331	S058T	24	1	38.4	2	3.30	2	1
332	S060T	22	3	39.0	2	3.20	1	2
333	S065T	21	2	39.9	1	3.60	1	1
334	S067T	24	2	37.9	1	3.60	1	1
335	S071T	29	2	38.1	1	4.40	1	2
336	S073T	24	3	35.7	2	2.85	2	1
337	S076T	18	1	40.3	1	3.62	1	1
338	S077T	30	2	39.0	2	3.65	1	2
339	S080T	18	1	34.4	1	3.72	2	1
340	S081T	24	1	41.3	1	3.24	1	1
341	S088T	23	2	37.7	1	3.00	1	1
342	S092T	23	1	39.1	1	3.90	1	1
343	S098T	23	3	36.4	2	3.00	1	2
344	S101T	24	1	40.4	1	2.83	1	1
345	S108T	26	2	38.9	2	3.82	1	1
346	S109T	32	5	40.0	2	3.51	1	2
347	S121T	22	2	38.9	1	2.82	1	1
348	S131T	26	2	40.4	1	3.73	1	1
349	S133T	27	2	38.0	1	3.80	2	2
350	S134T	22	2	36.3	1	3.82	1	2

No	ID	Mage	Gr	Bage	Gender	Bmass	Delmeth	Onset
351	T002P	24	1	38.0	2	3.20	1	2
352	T003P	30	3	38.1	2	3.75	1	1
353	T004P	29	3	37.3	1	2.15	1	1
354	T002T	21	1	41.0	2	3.01	2	1
355	T003T	32	3	37.3	2	2.95	2	1
356	T010T	25	2	36.9	2	3.00	2	2
357	T012T	24	1	39.7	1	3.62	2	2
358	T017T	19	1	39.6	2	3.70	1	2
359	T022T	23	1	39.4	2	3.28	2	2
360	T027T	21	1	39.0	1	4.20	1	1
361	U002T	28	5	38.7	1	3.15	1	1
362	V003P	24	1	37.1	1	3.10	1	2
363	V011P	19	1	39.4	2	3.58	1	1
364	V013P	22	1	39.6	1	3.96	2	2
365	V003T	28	3	38.6	1	3.52	1	1
366	V005T	20	1	40.7	1	3.70	1	1
367	V007T	29	3	38.6	2	3.42	1	1
368	V013T	18	1	42.0	2	3.00	1	2
369	V014T	36	3	38.0	1	3.80	1	1
370	V016T	29	3	41.3	2	3.75	1	2
371	V025T	20	1	40.3	2	3.74	1	1
372	V027T	23	1	37.4	2	3.50	1	2
373	V029T	38	4	34.0	2	2.75	1	1
374	V030T	22	2	40.3	2	3.18	1	1
375	V034T	28	2	38.3	1	3.75	2	1
376	V039T	24	3	40.0	2	3.36	1	1
377	V045T	24	3	39.1	1	4.00	2	2
378	V046T	21	2	39.4	2	2.30	1	1
379	V051T	25	1	42.7	2	3.38	1	1
380	V057T	22	4	38.1	1	3.10	1	1
381	V061T	19	1	40.6	1	3.58	2	1
382	V062T	25	2	38.6	1	3.40	2	1
383	V065T	20	1	38.9	1	3.36	2	1
384	V066T	28	2	40.6	1	3.32	1	2
385	V078T	21	2	39.4	2	3.10	1	1
386	V079T	24	1	38.6	2	3.86	1	2
387	V082T	27	4	38.9	2	3.77	1	1
388	V084T	20	1	40.0	2	3.40	2	1
389	V089T	31	4	39.3	1	3.50	1	1
390	V090T	26	3	35.0	1	2.30	1	2
391	V092T	23	4	36.0	2	2.50	2	1
392	V103T	21	1	39.3	2	3.02	1	1
393	V104T	18	4	38.0	2	2.71	1	1
394	V105T	23	1	34.4	1	2.41	2	2
395	V108T	25	3	35.6	1	2.98	1	1

No	ID	Mage	Gr	Bage	Gender	Bmass	Delmeth	Onset
396	V109T	21	1	38.4	2	3.00	1	2
397	V113T	20	2	42.0	1	4.56	1	2
398	V115T	28	3	41.6	1	3.74	2	1
399	V123T	22	1	40.1	2	3.25	2	1
400	V124T	28	4	37.1	2	2.86	1	1
401	V132T	24	2	38.6	1	3.50	1	1
402	V140T	38	1	37.6	2	3.35	2	1
403	V149T	24	5	39.6	1	3.34	1	1
404	V150T	23	4	39.0	1	2.90	1	2
405	V152T	22	1	41.3	1	3.35	2	1
406	V154T	19	1	35.1	1	2.50	1	1
407	V155T	31	2	38.4	1	3.10	1	1
408	V173T	16	1	35.4	2	3.62	1	1
409	V180T	33	4	38.9	1	2.75	1	1
410	V181T	28	2	42.0	1	4.50	1	1
411	V184T	27	2	38.1	1	3.18	1	2
412	V185T	21	2	39.4	1	3.10	1	1
413	V187T	20	1	38.0	1	3.42	1	2
414	V188T	21	1	38.3	1	3.22	1	2
415	V189T	21	1	40.4	1	3.50	2	1
416	V192T	39	9	38.9	2	3.34	1	1
417	V193T	29	4	35.6	1	4.58	2	2
418	V194T	32	6	40.0	2	3.74	1	2
419	V195T	26	2	37.6	2	3.50	1	2
420	V202T	26	2	38.9	1	4.00	1	1
421	V207T	24	2	39.4	1	3.38	1	2
422	V209T	21	1	37.9	2	2.90	1	1
423	V213T	20	1	38.3	1	2.76	2	1
424	V216T	20	1	41.9	1	3.62	2	2
425	V217T	23	1	38.0	2	3.42	1	1
426	V222T	20	2	40.6	2	3.05	1	1
427	V223T	19	1	41.7	1	3.70	1	1
428	W004P	28	1	40.7	2	4.06	1	2
429	W007T	24	1	39.1	1	4.30	2	1
430	W014T	24	1	40.6	1	3.70	2	1
431	W015T	22	3	38.7	2	2.61	2	1
432	W022T	26	2	42.6	1	3.40	1	2
433	W027T	28	4	40.4	1	3.80	2	2
434	W028T	29	3	37.7	2	2.75	1	1
435	W035T	25	1	39.0	1	3.80	1	1
436	W038T	36	4	40.3	1	3.50	1	2
437	W044T	20	1	39.4	2	3.60	1	1
438	W050T	21	1	40.0	2	3.70	1	2
439	W051T	32	1	37.3	1	3.30	2	1
440	Z001T	20	1	38.3	2	3.90	1	1

Appendix A

Data Set 4

A competitive ninety-seat steakhouse is well supported by the local community. For two months data was gathered on a daily basis and the days were classified as 'normal days' and 'special days' when more than an average number of visitors was expected, for example weekends, holidays, the day preceding a public holiday, etcetera).

The variables in the data set are as follows:

Weekday	=	Day of the week.
Date	=	The date.
Clients	=	The number of clients that visited the steakhouse.
Turnover	=	The turnover for the day.
Special	=	Special or normal day 1 = Special day 0 = Normal day.

April 1991

Weekday	Date	Clients	Turnover (R)	Special
Monday	1 April 1991	112	2 254.87	1
Tuesday	2 April 1991	54	903.06	0
Wednesday	3 April 1991	70	862.04	0
Thursday	4 April 1991	65	910.02	0
Friday	5 April 1991	90	1 552.86	1
Saturday	6 April 1991	101	2 470.77	1
Sunday	7 April 1991	80	1 510.02	1
Monday	8 April 1991	41	669.08	0
Tuesday	9 April 1991	53	764.59	0
Wednesday	10 April 1991	68	819.09	0
Thursday	11 April 1991	32	509.55	0
Friday	12 April 1991	73	1 158.43	1
Saturday	13 April 1991	82	1 362.57	1
Sunday	14 April 1991	70	1 183.64	1
Monday	15 April 1991	62	1 049.69	0
Tuesday	16 April 1991	48	661.70	0
Wednesday	17 April 1991	50	702.65	0
Thursday	18 April 1991	42	671.58	0
Friday	19 April 1991	63	1 067.83	1
Saturday	20 April 1991	73	1 043.66	1
Sunday	21 April 1991	70	830.02	1
Monday	22 April 1991	42	567.54	0
Tuesday	23 April 1991	53	610.17	0
Wednesday	24 April 1991	72	1 088.07	0
Thursday	25 April 1991	89	1 690.00	0
Friday	26 April 1991	99	2 305.88	1
Saturday	27 April 1991	142	2 886.32	1
Sunday	28 April 1991	159	2 989.03	1
Monday	29 April 1991	75	1 829.51	0
Tuesday	30 April 1991	172	2 537.53	1

May 1991

Weekday	Date	Clients	Turnover (R)	Special
Wednesday	1 May 1991	142	1 945.08	1
Thursday	2 May 1991	83	1 085.25	0
Friday	3 May 1991	112	2 549.49	1
Saturday	4 May 1991	99	1 392.50	1
Sunday	5 May 1991	83	1 271.22	1
Monday	6 May 1991	35	705.01	0
Tuesday	7 May 1991	58	850.05	0
Wednesday	8 May 1991	111	1 959.86	1
Thursday	9 May 1991	87	1 201.92	1
Friday	10 May 1991	68	1 034.41	1
Saturday	11 May 1991	75	1 419.21	1
Sunday	12 May 1991	134	3 094.86	1
Monday	13 May 1991	51	821.60	0
Tuesday	14 May 1991	49	445.87	0
Wednesday	15 May 1991	43	818.53	0
Thursday	16 May 1991	47	926.41	0
Friday	17 May 1991	78	851.83	1
Saturday	18 May 1991	80	848.46	1
Sunday	19 May 1991	78	1 070.09	1
Monday	20 May 1991	53	684.15	0
Tuesday	21 May 1991	54	878.01	0
Wednesday	22 May 1991	63	1 198.23	0
Thursday	23 May 1991	72	990.45	0
Friday	24 May 1991	92	2 430.71	1
Saturday	25 May 1991	112	3 007.54	1
Sunday	26 May 1991	109	2 900.65	1
Monday	27 May 1991	83	2 005.88	0
Tuesday	28 May 1991	148	4 400.32	0
Wednesday	29 May 1991	113	2 578.90	1
Thursday	30 May 1991	155	3 696.45	1
Friday	31 May 1991	138	3 165.31	1

Appendix A

Data Set 5

A brewery has a problem with the amount of electricity consumed. The annual electricity consumption (in relation to the amount of beer brewed) is considerably higher than for comparable breweries. A project was launched to decrease the electricty consumption for the bookyear 1 April 1992 to 31 March 1993. The amount of beer produced weekly and the weekly electricity consumption were recorded for the 1992 and 1993 bookyears. The data was obtained from the MBL group PBG00193.

The variables are as follows:

Week	=	The week of the year.
kWh	=	The weekly electricity consumption (in Kwh).
hl	=	The amount of beer produced weekly (in hectolitres).
kWh/hl	=	The electricity consumption per hectolitre of beer produced.

1992 Bookyear			
Week	kWh	hl	kWh/hl
91 / 13	248 510	14 020	17.73
14	214 270	19 557	10.96
15	265 510	28 117	9.44
16	210 590	11 620	18.12
17	207 940	14 032	14.82
18	173 790	10 100	17.21
19	170 980	13 226	12.93
20	187 260	15 793	11.86
21	149 050	11 852	12.58
22	125 360	11 852	10.58
23	165 020	18 162	9.09
24	155 020	20 930	7.41
25	145 020	12 333	11.76
26	144 020	6 330	22.75
27	133 020	4 553	29.22
28	85 020	2 600	32.70
29	235 020	7 603	30.91
30	165 020	12 333	13.38
31	230 010	20 196	11.39
32	230 010	20 196	11.39
33	236 890	19 453	12.18
34	251 740	23 337	10.79
35	264 440	29 565	8.94
36	264 170	21 437	12.32
37	270 240	25 799	10.47
38	268 020	30 143	8.89
39	273 980	26 643	10.28
40	302 070	24 405	12.38
41	248 630	22 851	10.88
42	314 650	29 588	10.63
43	289 280	27 010	10.71
44	323 920	29 509	10.98
45	230 540	16 416	14.04
46	284 740	27 349	10.41
47	305 700	27 029	11.31
48	284 110	26 553	10.70
49	309 880	25 325	12.24
50	307 450	25 897	11.87
51	297 300	24 486	12.14
52	79 392	6 616	12.00
53	38 785	4 872	7.96
92 / 1	297 300	24 486	12.14
2	231 840	20 004	11.59
3	268 110	16 104	16.65
4	243 100	20 337	11.95
5	269 370	19 721	13.66
6	267 310	20 440	13.08
7	258 860	20 203	12.81
8	256 790	15 044	17.07
9	254 070	12 617	20.14
10	209 370	19 721	10.62
11	176 800	16 659	10.61
12	209 370	19 721	10.62

1993 Bookyear			
Week	kWh	hl	kWh/hl
92 / 13	135 690	7 191	18.87
14	221 500	12 364	17.91
15	203 630	14 585	13.96
16	226 650	15 778	14.36
17	187 520	15 775	11.89
18	197 920	14 619	13.54
19	221 440	19 301	11.47
20	225 520	21 741	10.37
21	244 560	15 732	15.55
22	175 040	8 686	20.15
23	212 340	15 977	13.29
24	193 838	14 372	13.49
25	172 030	12 485	13.78
26	191 083	16 953	11.27
27	200 240	17 494	11.45
28	204 280	18 148	11.26
29	204 920	17 367	11.80
30	206 050	19 931	10.34
31	203 850	19 997	10.19
32	160 280	12 891	12.43
33	200 550	20 070	9.99
34	199 440	19 839	10.05
35	192 510	18 098	10.64
36	200 430	16 098	12.45
37	211 630	21 734	9.74
38	211 980	19 426	10.91
39	203 310	19 994	10.17
40	217 880	20 424	10.67
41	210 010	20 140	10.43
42	221 470	20 473	10.82
43	257 230	28 370	9.07
44	248 250	27 603	8.99
45	252 150	23 808	10.59
46	256 520	24 807	10.34
47	261 460	25 565	10.23
48	269 810	27 792	9.71
49	276 250	33 766	8.18
50	310 300	37 409	8.29
51	299 870	34 238	8.76
52	200 880	18 987	10.58
53	209 370	19 992	10.47
93 / 1	278 650	27 116	10.28
2	242 590	19 360	12.53
3	227 570	18 828	12.09
4	211 200	27 003	7.82
5	212 470	21 820	9.74
6	222 460	18 090	12.30
7	223 200	17 494	12.76
8	194 580	10 362	18.78
9	219 660	9 078	24.20
10	256 300	24 610	10.41
11	237 600	24 366	9.75
12	245 840	25 220	9.75

Appendix B *Statistical tables*

The tables in this appendix are a prerequisite for applying the statistical inference techniques that are discussed in this book.

Table B.1	Random numbers
Table B.2	Area under the normal curve
Table B.3	Critical values for the t distribution
Table B.4	Critical values for the χ^2 distribution
Table B.5	Critical values for the F distribution
Table B.6	Cumulative binomial distribution
Table B.7	Cumulative Poisson distribution
Table B.8	Critical values for the $r(n)$ distribution
Table B.9	The Z-transformation for the correlation coefficient
Table B.10	Critical values for the sign test
Table B.11	Critical values for the signed rank test
Table B.12	Critical values for the rank sum test
Table B.13	a values for $(I \times J)$ log linear analysis
Table B.14	Values for the inverse normal distribution

Table B.1 **Random numbers**

This table contains 10 000 random numbers. These numbers comply with all the requirements set for random numbers and can thus be regarded as a fully randomized array of numbers 0, 1, 2, ..., 9.

Row																									
1	83	20	37	15	72	22	21	78	47	52	14	51	94	38	68	69	47	58	04	45	29	85	48	58	79
2	55	06	80	64	34	71	41	32	97	37	73	05	04	33	04	96	16	42	50	72	91	30	86	66	17
3	19	54	68	57	50	27	10	08	16	26	41	89	75	46	88	83	12	20	30	51	03	73	77	24	61
4	42	49	74	33	75	24	22	79	32	12	55	23	63	66	06	78	83	60	61	46	73	84	85	52	04
5	17	87	57	75	85	31	11	81	00	94	86	28	26	23	87	68	11	76	65	80	64	19	47	83	39
6	90	73	51	07	27	30	68	26	75	42	33	35	11	88	03	83	17	00	29	82	94	50	42	39	41
7	67	37	66	97	75	98	54	02	23	60	80	96	13	08	79	28	48	01	72	28	26	07	95	83	89
8	76	19	85	46	62	46	64	30	58	35	62	75	25	68	52	53	64	28	49	57	66	52	70	54	74
9	97	76	37	48	02	83	03	63	59	78	55	25	54	22	99	44	39	35	84	50	73	51	68	77	81
10	55	24	37	37	72	35	96	54	05	45	26	00	71	29	15	03	85	34	15	27	79	77	49	03	64
11	12	29	63	58	84	75	61	53	11	09	69	69	11	16	62	31	48	82	73	21	82	10	70	87	56
12	40	74	22	02	30	93	56	33	50	94	36	37	34	26	28	23	36	31	73	73	57	55	92	43	03
13	28	23	78	80	23	98	97	16	99	86	94	69	93	93	70	67	68	53	11	04	04	18	05	22	91
14	18	11	67	73	29	59	78	37	66	62	66	71	08	77	20	98	80	13	08	21	44	68	66	60	15
15	79	36	14	75	23	41	74	46	91	00	12	78	89	97	21	01	40	08	31	96	92	75	06	64	08
16	89	02	33	69	91	95	70	70	22	65	40	85	89	96	54	05	45	44	72	34	23	32	60	43	08
17	52	83	17	02	17	85	35	75	34	75	21	38	18	83	18	48	55	27	66	63	36	33	91	01	87
18	49	93	37	01	71	80	38	09	76	30	64	04	23	28	33	86	43	28	39	29	56	35	60	91	47
19	19	82	36	96	12	53	94	00	21	77	81	03	36	83	31	69	07	01	27	71	54	01	27	38	83
20	01	02	33	71	85	69	39	89	92	42	93	08	86	45	99	19	59	56	11	01	70	16	66	07	28
21	70	23	28	73	71	33	44	28	15	69	24	45	62	11	68	35	63	23	13	06	61	52	45	61	74
22	02	99	40	01	23	37	36	41	82	55	72	42	91	06	35	69	09	32	23	23	40	18	58	78	76
23	36	61	46	37	95	73	89	20	50	77	44	60	30	83	47	72	26	84	00	82	34	15	26	26	23
24	28	68	39	42	28	36	85	39	29	10	22	62	75	29	58	30	14	78	99	99	28	14	27	45	51
25	23	95	60	71	84	24	49	90	38	42	10	95	81	95	19	17	39	68	45	90	46	20	03	25	83
26	92	14	05	83	10	38	71	74	92	04	26	93	75	07	29	01	10	21	74	86	91	04	51	76	40
27	37	15	22	86	98	54	72	10	96	29	99	51	81	60	27	46	36	30	30	68	67	92	26	05	84
28	96	84	03	37	18	07	76	06	69	67	36	43	65	98	31	48	18	96	00	13	91	25	37	69	66
29	13	07	59	57	60	30	63	65	80	63	25	46	71	86	92	19	84	71	89	49	00	56	09	07	08
30	95	63	25	05	08	32	08	71	92	17	51	10	02	06	62	77	76	07	75	85	30	33	85	11	73
31	29	81	10	89	59	20	66	52	04	19	60	65	90	57	89	40	42	15	92	94	43	04	47	16	07
32	49	81	91	00	08	95	58	17	46	79	68	56	12	26	69	28	94	09	74	70	57	51	01	51	69
33	33	78	85	24	00	90	71	75	17	58	49	31	96	04	71	66	42	32	15	05	93	06	36	23	89
34	77	02	91	18	09	69	40	05	06	76	73	50	39	11	92	04	70	93	51	96	57	16	13	13	07
35	36	76	64	09	38	54	65	02	33	62	07	07	22	24	98	05	87	00	48	05	21	25	43	52	66
36	83	31	69	00	66	14	88	09	89	21	36	64	73	10	78	78	82	56	77	58	58	60	63	57	64
37	04	89	25	40	49	78	97	34	31	91	83	93	41	68	46	07	00	40	14	09	14	39	40	44	25
38	28	92	55	43	44	85	74	38	94	96	28	65	23	49	79	93	56	92	43	65	64	20	46	67	47
39	04	18	42	87	45	19	49	65	57	43	08	00	06	04	44	68	52	16	27	07	21	38	20	43	30
40	79	50	22	49	76	10	60	08	60	67	69	73	22	86	61	36	06	98	69	32	91	39	05	85	49
41	11	18	15	65	05	34	63	70	70	99	78	69	49	66	85	87	91	24	05	89	23	83	01	74	02
42	04	73	57	27	95	44	38	34	04	11	69	09	53	20	59	59	49	70	88	26	23	70	99	48	19
43	55	15	18	85	46	80	76	11	89	87	19	58	44	72	19	90	31	43	06	67	70	10	72	35	83
44	25	62	90	52	56	08	32	91	29	46	04	39	00	90	65	74	77	63	92	58	11	67	27	96	32
45	91	44	73	08	52	43	46	56	36	20	70	11	96	34	27	45	52	74	59	07	15	53	91	65	16
46	07	14	89	71	48	60	16	23	00	97	32	78	87	68	27	87	93	49	82	01	69	84	89	10	37
47	68	42	19	79	42	25	65	38	58	01	61	57	89	77	82	13	13	87	18	99	01	94	97	92	75
48	59	12	53	02	04	36	97	09	34	01	46	86	83	20	41	75	04	96	47	48	73	15	11	46	71
49	61	77	16	15	83	14	34	15	11	54	77	26	69	48	11	55	13	74	16	87	78	15	49	37	08
50	19	35	98	88	38	08	81	59	64	20	71	76	63	30	51	61	39	26	29	84	15	66	57	43	94

Table B.1 **Random numbers** (continued)

Row																										
1	69	35	58	61	78	25	51	01	69	13	17	19	20	37	21	00	19	15	07	16	45	63	58	25	11
2	13	17	27	23	57	87	75	63	35	94	22	01	96	87	63	79	60	64	19	34	27	47	36	22	66
3	64	96	34	84	30	01	17	73	44	72	19	64	17	52	67	29	84	09	53	62	78	69	23	89	78
4	09	54	98	23	58	07	91	65	18	10	28	45	33	85	43	93	82	81	45	01	60	07	27	96	34
5	82	36	56	29	42	54	52	81	62	89	27	79	43	86	29	85	35	82	61	11	76	36	46	13	91
6	37	00	10	37	54	37	95	54	35	70	89	56	26	49	83	87	98	74	49	27	80	44	38	77	00
7	32	27	73	99	15	93	36	35	41	30	00	98	73	33	85	23	58	98	54	59	30	97	47	34	27
8	46	43	91	39	89	40	40	32	44	95	34	94	90	41	00	11	16	28	29	62	21	78	43	94	81
9	57	45	18	04	22	79	76	79	29	71	31	10	00	14	57	09	27	87	33	28	80	33	10	18	04
10	40	21	59	89	43	78	99	73	65	69	87	25	56	02	28	52	75	35	33	98	00	72	41	52	37
11	72	12	73	96	25	93	62	45	49	51	06	53	24	06	99	94	82	15	22	90	64	35	24	97	46
12	67	66	80	59	70	94	62	95	94	10	71	02	84	09	96	87	31	99	87	24	95	74	23	58	50
13	78	70	84	28	71	61	23	63	37	03	78	09	75	27	45	91	98	70	23	27	74	45	84	69	13
14	09	15	22	38	77	20	71	93	98	87	94	94	91	64	47	20	36	99	90	06	82	01	80	28	61
15	98	34	80	43	07	29	19	78	59	40	18	71	19	92	21	20	52	88	14	24	24	73	66	37	81
16	76	44	70	15	31	33	29	00	70	22	32	51	18	21	52	76	18	62	97	51	05	09	61	44	49
17	00	68	79	02	28	39	94	69	07	09	32	07	21	38	94	59	79	49	55	50	48	53	17	30	05
18	99	20	02	04	28	99	30	17	18	62	84	68	72	34	29	27	85	37	92	84	92	31	47	24	87
19	50	17	50	22	91	04	03	59	74	90	07	49	91	78	53	29	23	44	66	28	50	59	86	05	83
20	36	09	54	15	43	47	33	57	83	38	51	00	13	82	92	03	84	46	93	22	62	89	05	98	12
21	13	72	17	87	44	52	56	35	78	75	83	14	64	59	76	92	81	82	08	88	45	25	28	45	30
22	94	80	46	51	57	59	90	21	30	00	50	47	97	49	35	82	11	24	71	73	94	06	84	72	32
23	39	67	15	51	57	94	93	54	15	30	75	95	95	94	31	03	79	80	57	86	68	10	62	03	91
24	97	72	28	67	32	46	08	68	54	33	14	42	73	47	79	69	10	62	99	79	70	67	73	52	68
25	53	41	89	04	22	17	87	21	33	69	48	83	47	86	03	99	56	84	87	65	01	50	66	14	66
26	51	59	92	34	45	44	52	79	09	09	94	01	38	54	81	19	66	58	25	75	91	10	83	85	33
27	67	73	46	57	31	25	13	11	97	67	94	02	28	39	40	14	49	48	79	26	97	43	34	05	83
28	10	23	39	87	65	68	33	71	66	39	84	18	74	52	56	90	44	31	12	12	50	64	50	14	00
29	88	25	98	17	73	86	85	97	16	69	78	21	10	15	27	47	30	51	16	44	68	24	99	07	88
30	25	14	38	97	39	36	20	44	72	53	27	89	19	95	83	05	46	77	52	34	05	97	94	88	24
31	88	90	64	01	96	11	57	62	26	86	05	83	92	85	39	49	23	31	41	64	50	75	53	09	39
32	61	37	23	27	05	24	90	43	72	99	29	03	73	77	63	79	41	63	14	86	09	42	91	23	01
33	53	12	77	00	13	19	29	98	53	59	23	06	71	66	61	32	06	21	99	91	10	00	09	68	77
34	00	73	84	95	97	09	11	62	71	10	42	21	78	71	97	52	67	22	09	80	65	89	85	31	19
35	79	14	80	94	01	32	93	12	81	63	47	85	47	35	27	44	71	67	35	78	17	61	51	06	20
36	86	89	63	49	73	35	56	99	15	07	66	58	24	01	61	99	76	87	24	96	14	08	76	76	65
37	95	99	11	49	49	95	84	48	82	05	32	83	95	55	29	83	95	28	22	87	20	70	15	67	01
38	30	29	91	15	60	45	83	05	25	52	10	07	10	73	94	81	55	19	77	59	24	94	43	82	31
39	50	84	15	38	03	10	80	93	58	54	68	57	59	79	01	10	56	95	48	89	06	13	29	88	24
40	90	99	69	11	83	27	12	36	24	54	69	79	69	73	80	00	33	68	51	74	75	57	50	38	78
41	71	18	95	37	73	46	78	86	53	15	51	44	64	97	95	53	29	30	76	21	16	86	05	78	16
42	25	56	68	25	58	03	66	36	08	85	82	79	20	10	95	49	61	16	15	21	57	28	64	39	61
43	02	40	68	26	54	80	01	64	50	05	52	23	39	04	47	40	02	19	24	20	62	27	70	16	12
44	48	12	87	31	21	39	22	35	14	78	35	62	54	29	46	90	92	82	09	30	01	66	14	66	69
45	69	08	53	53	64	88	59	25	56	46	75	10	37	13	94	00	25	79	32	92	70	90	14	30	71
46	83	88	83	85	93	67	66	57	50	22	82	17	96	21	75	85	27	80	17	39	71	67	69	08	60
47	24	00	95	21	08	99	96	62	65	41	69	08	88	25	13	50	51	95	37	68	57	58	07	80	38
48	53	73	46	64	14	96	28	03	85	89	67	65	70	05	44	72	35	73	96	72	11	72	68	58	82
49	61	88	71	39	35	82	62	56	20	16	86	50	45	83	13	94	22	41	79	96	34	72	98	72	41
50	54	51	21	24	35	04	59	25	74	30	45	92	75	05	97	73	38	96	08	73	15	27	37	37	28

Table B.1 **Random numbers** (continued)

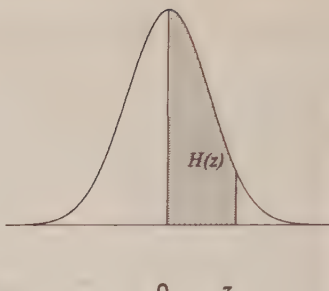
Row																									
1	05	82	59	43	12	46	05	83	00	85	31	59	83	40	40	20	27	58	13	91	39	01	04	61	89
2	62	89	27	39	36	49	72	84	41	25	64	69	20	44	22	99	90	74	18	17	64	30	85	20	64
3	09	20	65	80	20	90	84	91	45	69	92	54	80	47	93	87	98	13	94	47	62	32	02	11	20
4	91	50	44	01	00	09	11	08	98	30	94	68	49	55	54	74	36	36	43	88	21	32	98	18	57
5	27	01	13	98	92	88	18	80	23	43	86	98	53	43	91	85	05	77	34	64	15	49	69	54	06
6	02	16	81	03	59	26	25	77	39	88	98	50	67	68	47	81	97	85	13	93	76	88	76	63	05
7	83	21	22	96	41	80	45	98	74	25	68	69	29	03	67	89	82	00	41	47	63	78	76	17	71
8	08	83	00	46	50	55	71	29	31	32	03	86	90	85	48	35	18	59	04	67	90	72	45	45	54
9	69	70	08	40	80	21	53	75	59	23	92	39	62	58	01	60	58	45	14	50	38	61	28	01	61
10	17	86	72	30	28	32	96	54	39	17	86	92	10	38	39	57	44	49	53	28	88	30	68	40	68
11	53	54	76	59	01	75	95	16	18	30	87	65	01	56	86	09	73	43	34	10	14	86	05	05	94
12	49	60	82	05	43	45	22	01	80	04	78	53	61	66	36	14	87	34	77	89	99	29	02	59	33
13	56	21	68	81	78	88	39	20	22	59	65	54	82	48	16	87	51	96	85	82	07	40	42	31	23
14	45	90	64	19	54	55	11	20	98	69	46	72	79	81	77	17	17	33	33	24	41	02	38	84	31
15	31	47	37	44	12	34	88	15	76	06	21	37	55	73	20	36	43	64	41	01	69	86	70	53	63
16	50	15	29	95	78	75	04	04	20	67	08	96	36	94	09	28	66	36	54	11	40	39	14	66	27
17	59	31	67	76	61	25	33	17	42	75	99	46	87	02	11	61	37	22	81	81	21	59	04	05	94
18	18	03	59	78	22	66	86	27	11	99	67	53	08	64	80	23	16	58	11	80	20	90	07	05	69
19	14	47	55	76	56	70	00	15	83	37	28	71	63	20	78	04	96	33	31	41	36	56	09	13	28
20	26	13	41	73	86	13	78	11	31	86	40	92	84	08	98	99	26	17	14	08	03	59	11	86	20
21	15	48	08	82	78	86	96	10	30	68	39	56	83	40	20	87	78	70	72	04	50	99	29	85	10
22	34	48	72	72	73	86	33	65	83	07	58	10	88	59	26	37	66	41	68	81	62	66	28	92	43
23	65	69	29	96	92	82	14	52	33	06	55	34	34	39	63	15	49	35	96	50	96	26	46	08	40
24	80	57	63	54	67	88	74	91	46	19	76	72	31	48	54	06	50	77	23	54	84	31	47	43	70
25	05	19	50	15	10	46	57	33	35	43	32	26	28	69	87	42	56	19	84	74	66	30	97	88	57
26	46	82	58	38	69	60	45	30	87	20	99	49	54	37	62	39	85	28	20	42	19	85	31	71	67
27	37	56	45	19	41	81	01	64	56	36	33	98	74	39	86	04	53	89	97	08	75	06	01	59	92
28	61	87	51	22	82	64	80	59	68	96	17	21	16	06	54	01	76	42	26	71	68	61	37	44	30
29	43	09	34	04	22	74	50	40	59	38	48	42	90	07	82	26	67	56	28	30	86	10	18	80	60
30	08	61	37	97	92	99	04	86	18	13	87	49	03	66	49	21	05	95	57	49	00	99	29	97	93
31	64	34	51	15	75	50	96	34	18	62	57	71	76	60	73	67	07	58	27	86	93	09	51	45	85
32	03	80	28	68	55	12	40	15	33	71	83	99	47	57	53	11	16	29	84	66	83	14	44	87	86
33	07	66	93	41	79	03	53	88	96	74	55	26	12	82	54	73	03	75	76	67	86	71	66	42	14
34	19	56	34	12	33	15	53	20	73	71	47	44	72	28	30	70	00	06	74	95	30	00	19	91	22
35	61	31	26	88	76	04	58	26	18	93	98	86	34	23	49	89	18	11	86	27	53	00	53	13	33
36	86	46	76	51	40	15	84	80	00	65	17	87	05	61	12	04	51	40	63	38	72	56	64	23	22
37	79	29	04	49	48	03	68	07	73	14	83	76	97	12	80	32	03	58	32	44	12	69	41	40	93
38	39	91	64	66	27	53	09	68	69	55	69	06	15	07	00	16	38	61	36	59	41	94	23	00	08
39	18	69	34	50	24	86	13	87	38	80	17	52	20	44	72	90	07	93	70	90	93	01	23	50	41
40	72	31	56	06	13	40	42	30	91	20	97	74	22	82	57	27	11	52	63	48	05	84	87	88	06
41	95	71	75	06	78	52	62	30	12	88	32	03	71	57	22	22	17	84	75	21	45	19	43	86	86
42	94	75	80	44	52	98	54	85	87	95	50	68	06	43	47	14	13	29	73	77	30	50	43	30	49
43	61	64	25	07	29	13	08	66	61	71	95	46	85	75	02	71	00	81	62	64	18	97	54	68	88
44	31	22	03	05	10	37	37	72	46	14	24	97	43	37	52	42	27	93	21	25	52	36	36	60	67
45	88	36	06	14	67	38	48	55	20	41	78	75	77	58	53	11	18	10	09	48	62	82	62	38	76
46	81	17	66	58	06	41	94	78	33	99	07	35	10	72	73	06	82	67	54	23	33	51	18	80	38
47	31	70	02	98	29	45	04	23	94	16	73	76	69	77	59	85	40	22	36	08	09	61	30	87	92
48	84	80	17	22	82	60	79	58	38	97	61	98	54	66	24	66	43	05	83	56	27	52	56	85	25
49	36	42	84	95	76	65	27	05	47	49	77	63	74	80	28	30	36	35	97	85	52	55	42	29	94
50	43	93	58	45	07	35	16	87	78	15	52	43	73	77	23	59	65	40	54	52	60	07	12	91	50

Table B.1 **Random numbers** (continued)

Row																									
1	38	53	41	52	77	97	74	90	36	41	22	96	35	88	47	56	15	31	10	96	96	91	11	66	23
2	14	88	27	32	18	21	07	08	75	64	07	42	20	23	09	35	63	35	09	43	63	73	42	62	82
3	33	63	17	37	35	75	59	24	30	32	82	37	88	02	39	89	71	69	98	69	46	99	39	37	06
4	76	85	31	10	40	41	80	52	29	11	81	16	77	32	95	97	83	76	66	26	67	10	87	72	09
5	88	01	30	82	14	46	86	66	01	57	93	61	47	52	34	17	44	69	10	01	27	88	99	72	10
6	14	00	83	19	69	27	50	66	40	88	90	23	36	99	19	57	07	02	26	54	08	58	99	65	40
7	54	54	47	23	42	33	26	04	30	85	28	31	70	21	78	53	57	53	47	52	13	81	88	31	84
8	85	52	76	66	33	63	39	10	18	74	99	33	70	17	18	03	80	97	63	06	40	42	32	83	38
9	64	32	83	92	95	63	56	09	14	71	98	23	69	61	25	39	40	02	17	78	10	33	68	33	16
10	90	70	12	77	15	40	12	73	88	30	10	23	37	18	56	93	98	99	24	29	30	88	64	59	91
11	95	95	61	10	17	01	66	60	72	87	02	32	87	37	13	34	42	81	56	92	44	56	21	16	83
12	56	32	31	30	66	15	24	56	17	40	29	94	00	12	33	73	64	18	25	57	24	29	10	63	56
13	90	93	22	32	34	01	85	87	87	63	66	14	63	79	57	38	92	05	07	74	46	03	08	34	77
14	22	68	40	08	23	41	52	62	62	32	86	83	94	22	83	68	53	92	91	02	26	27	52	95	83
15	01	47	89	99	92	81	42	74	55	12	39	56	07	18	42	44	99	03	84	21	09	58	19	51	91
16	73	96	84	59	04	62	00	32	58	39	21	40	58	31	03	87	17	49	20	42	69	37	50	72	90
17	11	02	17	48	69	10	45	17	64	03	71	52	28	69	85	47	21	09	62	99	95	18	16	02	74
18	63	63	03	06	86	79	22	60	58	49	68	40	33	71	80	77	60	66	96	73	64	36	59	83	14
19	81	89	60	01	39	12	53	76	96	61	30	75	95	76	03	06	36	58	06	32	54	67	18	69	21
20	96	47	18	94	90	05	60	90	70	89	56	32	50	50	83	20	87	19	21	75	11	00	20	66	27
21	77	60	77	58	53	03	38	51	17	32	98	11	08	95	82	06	65	20	41	96	06	59	05	78	94
22	04	21	59	29	44	30	44	71	60	74	98	19	10	00	89	55	81	74	67	96	85	61	22	66	92
23	89	40	97	47	05	75	10	15	16	93	37	88	52	40	09	07	17	51	75	84	93	19	60	47	16
24	64	23	21	50	59	30	63	62	38	97	70	29	11	80	96	94	00	50	67	67	35	03	91	19	20
25	47	09	35	19	65	71	61	31	30	73	67	66	40	93	85	46	50	23	36	20	70	72	22	49	80
26	58	70	04	80	52	12	54	35	26	68	33	16	90	06	61	48	80	98	92	13	66	47	31	07	87
27	88	25	67	75	85	86	06	35	76	40	48	92	54	04	97	89	73	43	19	18	39	34	16	84	60
28	70	04	01	40	91	43	94	04	42	52	89	25	54	34	48	57	43	89	49	62	86	52	34	97	99
29	35	69	24	36	56	42	04	66	74	70	41	84	28	74	51	94	96	55	74	30	80	00	60	44	46
30	85	21	62	70	72	13	82	04	33	07	95	02	24	97	61	06	00	86	06	76	97	32	28	66	13
31	60	69	86	93	29	13	90	58	54	91	11	81	56	50	53	46	37	07	09	97	86	94	96	68	58
32	79	12	40	58	92	38	30	17	40	20	47	35	72	79	54	43	83	40	41	16	90	95	32	05	95
33	94	97	08	82	62	37	09	94	81	95	51	51	88	54	68	07	61	18	98	71	95	84	47	21	99
34	03	14	30	66	77	01	07	23	58	32	56	29	05	25	60	83	29	12	94	60	43	90	09	06	57
35	47	76	39	25	81	16	72	79	63	29	14	47	30	85	24	34	65	63	62	07	01	19	63	42	74
36	55	27	45	90	71	55	39	16	51	60	71	48	97	14	86	01	62	30	09	73	27	88	76	82	78
37	79	34	98	20	16	98	48	91	03	76	84	21	97	94	22	56	53	85	80	16	15	78	97	78	21
38	31	38	71	89	90	19	49	79	27	55	91	25	30	20	43	69	70	93	40	14	68	68	32	04	17
39	50	49	64	74	87	31	51	14	67	16	17	71	61	58	31	43	41	51	15	47	72	66	94	81	15
40	51	37	41	68	77	81	54	93	70	81	93	59	88	83	80	88	15	96	42	07	17	66	33	05	16
41	13	92	62	98	41	56	65	74	39	46	93	55	86	17	82	55	10	16	91	31	71	92	03	72	10
42	01	91	50	71	64	56	35	82	51	11	31	72	72	35	78	77	56	33	59	42	20	19	68	76	37
43	96	06	41	93	42	74	73	62	26	66	49	63	48	43	00	11	77	34	47	21	96	41	75	26	72
44	98	40	95	36	27	86	92	38	97	80	95	49	82	35	80	79	16	51	81	02	84	53	38	54	39
45	38	28	14	50	16	00	72	79	63	48	23	65	62	94	25	45	28	16	28	92	77	94	38	62	98
46	53	58	26	11	17	63	64	61	39	13	70	01	03	75	43	79	55	43	57	08	95	62	64	98	38
47	79	87	44	92	78	04	43	87	70	03	22	77	39	91	59	33	63	54	13	00	09	87	33	38	52
48	35	89	85	45	93	09	74	73	42	74	64	39	13	98	94	04	03	24	49	53	49	49	04	10	01
49	11	62	82	96	22	17	84	39	48	15	74	51	80	42	27	97	67	92	97	13	11	00	01	45	70
50	92	78	73	32	88	03	31	01	80	44	02	15	98	94	37	89	97	36	56	98	72	55	60	22	85

Table B.2 Area under the normal curve

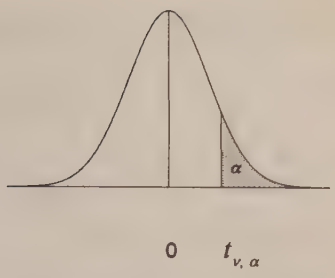
The table contains the values for $H(z)$ for positive values of z with $H(z) = P(0 < Z < z)$.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41309	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670
2.0	0.47725	0.47778	0.47831	0.47882	0.47932	0.47982	0.48030	0.48077	0.48124	0.48169
2.1	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461	0.48500	0.48537	0.48574
2.2	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.48840	0.48870	0.48899
2.3	0.48928	0.48956	0.48983	0.49010	0.49036	0.49061	0.49086	0.49111	0.49134	0.49158
2.4	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305	0.49324	0.49343	0.49361
2.5	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477	0.49492	0.49506	0.49520
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.9	0.49813	0.49819	0.49825	0.49831	0.49836	0.49841	0.49846	0.49851	0.49856	0.49861
3.0	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889	0.49893	0.49896	0.49900
3.1	0.49903	0.49906	0.49910	0.49913	0.49916	0.49918	0.49921	0.49924	0.49926	0.49929
3.2	0.49931	0.49934	0.49936	0.49938	0.49940	0.49942	0.49944	0.49946	0.49948	0.49950
3.3	0.49952	0.49953	0.49955	0.49957	0.49958	0.49960	0.49961	0.49962	0.49964	0.49965
3.4	0.49966	0.49968	0.49969	0.49970	0.49971	0.49972	0.49973	0.49974	0.49975	0.49976
3.5	0.49977	0.49978	0.49978	0.49979	0.49980	0.49981	0.49981	0.49982	0.49983	0.49983
3.6	0.49984	0.49985	0.49985	0.49986	0.49986	0.49987	0.49987	0.49988	0.49988	0.49989
3.7	0.49989	0.49990	0.49990	0.49990	0.49991	0.49991	0.49992	0.49992	0.49992	0.49992
3.8	0.49993	0.49993	0.49993	0.49994	0.49994	0.49994	0.49994	0.49995	0.49995	0.49995
3.9	0.49995	0.49995	0.49996	0.49996	0.49996	0.49996	0.49996	0.49996	0.49997	0.49997

Table B.3 Critical values of the t distribution

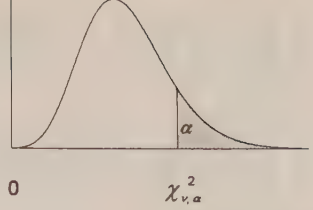
The table contains the $t_{\nu,\alpha}$ values for the given values of ν and α with $P(T_\nu \geq t_{\nu,\alpha}) = \alpha$.



ν	α							
	0.4	0.3	0.2	0.1	0.05	0.025	0.010	0.005
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.656
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750
35	0.255	0.529	0.852	1.306	1.690	2.030	2.438	2.724
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704
50	0.255	0.528	0.849	1.299	1.676	2.009	2.403	2.678
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660
70	0.254	0.527	0.847	1.294	1.667	1.994	2.381	2.648
80	0.254	0.526	0.846	1.292	1.664	1.990	2.374	2.639
90	0.254	0.526	0.846	1.291	1.662	1.987	2.368	2.632
100	0.254	0.526	0.845	1.290	1.660	1.984	2.364	2.626
∞	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576

Table B.4 Critical values of the χ^2 distribution

The table contains the $\chi^2_{\nu,\alpha}$ values for the given values of ν and α with $P(\chi_\nu^2 \geq \chi^2_{\nu,\alpha}) = \alpha$.



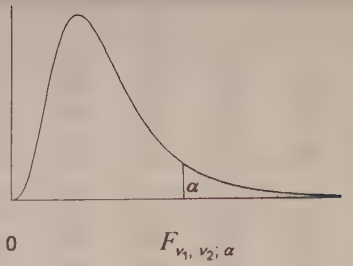
ν	α																	
	0.995	0.990	0.975	0.950	0.900	0.800	0.700	0.600	0.500	0.400	0.300	0.200	0.100	0.050	0.025	0.010	0.005	0.001
1	0.000	0.000	0.001	0.004	0.016	0.064	0.148	0.275	0.455	0.708	1.074	1.642	2.706	3.841	5.024	6.635	7.879	10.827
2	0.010	0.020	0.051	0.103	0.211	0.446	0.713	1.022	1.386	1.833	2.408	3.219	4.605	5.991	7.378	9.210	10.597	13.815
3	0.072	0.115	0.216	0.352	0.584	1.005	1.424	1.869	2.366	2.946	3.665	4.642	6.251	7.815	9.348	11.345	12.838	16.266
4	0.207	0.297	0.484	0.711	1.064	1.649	2.195	2.753	3.357	4.045	4.878	5.989	7.779	9.488	11.143	13.277	14.860	18.466
5	0.412	0.554	0.831	1.145	1.610	2.343	3.000	3.656	4.351	5.132	6.064	7.289	9.236	11.070	12.832	15.086	16.750	20.515
6	0.676	0.872	1.237	1.635	2.204	3.070	3.828	4.570	5.348	6.211	7.231	8.558	10.645	12.592	14.449	16.812	18.548	22.457
7	0.989	1.239	1.690	2.167	2.833	3.822	4.671	5.493	6.346	7.283	8.383	9.803	12.017	14.067	16.013	18.475	20.278	24.321
8	1.344	1.647	2.180	2.733	3.490	4.594	5.527	6.423	7.344	8.351	9.524	11.030	13.362	15.507	17.535	20.090	21.955	26.124
9	1.735	2.088	2.700	3.325	4.168	5.380	6.393	7.357	8.343	9.414	10.656	12.242	14.684	16.919	19.023	21.666	23.589	27.877
10	2.156	2.558	3.247	3.940	4.865	6.179	7.267	8.295	9.342	10.473	11.781	13.442	15.987	18.307	20.483	23.209	25.188	29.588
11	2.603	3.053	3.816	4.575	5.578	6.989	8.148	9.237	10.341	11.530	12.899	14.631	17.275	19.675	21.920	24.725	26.757	31.264
12	3.074	3.571	4.404	5.226	6.304	7.807	9.034	10.182	11.340	12.584	14.011	15.812	18.549	21.026	23.337	26.217	28.300	32.909
13	3.565	4.107	5.009	5.892	7.041	8.634	9.926	11.129	12.340	13.636	15.119	16.985	19.812	22.362	24.736	27.688	29.819	34.527
14	4.075	4.660	5.629	6.571	7.790	9.467	10.821	12.078	13.339	14.685	16.222	18.151	21.064	23.685	26.119	29.141	31.319	36.124
15	4.601	5.229	6.262	7.261	8.547	10.307	11.721	13.030	14.339	15.733	17.322	19.311	22.307	24.996	27.488	30.578	32.801	37.698
16	5.142	5.812	6.908	7.962	9.312	11.152	12.624	13.983	15.338	16.780	18.418	20.465	23.542	26.296	28.845	32.000	34.267	39.252
17	5.697	6.408	7.564	8.672	10.085	12.002	13.531	14.937	16.338	17.824	19.511	21.615	24.769	27.587	30.191	33.409	35.718	40.791
18	6.265	7.015	8.231	9.390	10.865	12.857	14.440	15.893	17.338	18.868	20.601	22.760	25.989	28.869	31.526	34.805	37.156	42.312
19	6.844	7.633	8.907	10.117	11.651	13.716	15.352	16.850	18.338	19.910	21.689	23.900	27.204	30.144	32.852	36.191	38.582	43.819
20	7.434	8.260	9.591	10.851	12.443	14.578	16.266	17.809	19.337	20.951	22.775	25.038	28.412	31.410	34.170	37.566	39.997	45.314
21	8.034	8.897	10.283	11.591	13.240	15.445	17.182	18.768	20.337	21.992	23.858	26.171	29.615	32.671	35.479	38.932	41.401	46.796
22	8.643	9.542	10.982	12.338	14.041	16.314	18.101	19.729	21.337	23.031	24.939	27.301	30.813	33.924	36.781	40.289	42.796	48.268
23	9.260	10.196	11.689	13.091	14.848	17.187	19.021	20.690	22.337	24.069	26.018	28.429	32.007	35.172	38.076	41.638	44.181	49.728
24	9.886	10.856	12.401	13.848	15.659	18.062	19.943	21.652	23.337	25.106	27.096	29.553	33.196	36.415	39.364	42.980	45.558	51.179
25	10.520	11.524	13.120	14.611	16.473	18.940	20.867	22.616	24.337	26.143	28.172	30.675	34.382	37.652	40.646	44.314	46.928	52.619
26	11.160	12.198	13.844	15.379	17.292	19.820	21.792	23.579	25.336	27.179	29.246	31.795	35.563	38.885	41.923	45.642	48.290	54.051
27	11.808	12.878	14.573	16.151	18.114	20.703	22.719	24.544	26.336	28.214	30.319	32.912	36.741	40.113	43.195	46.963	49.645	55.475
28	12.461	13.565	15.308	16.928	18.939	21.588	23.647	25.509	27.336	29.249	31.391	34.027	37.916	41.337	44.461	48.278	50.994	56.892
29	13.121	14.256	16.047	17.708	19.768	22.475	24.577	26.475	28.336	30.283	32.461	35.139	39.087	42.557	45.722	49.588	52.335	58.301
30	13.787	14.953	16.791	18.493	20.599	23.364	25.508	27.442	29.336	31.316	33.530	36.250	40.256	43.773	46.979	50.892	53.672	59.702
35	17.192	18.509	20.569	22.465	24.797	27.836	30.178	32.282	34.336	36.475	38.859	41.778	46.059	49.802	53.203	57.342	60.275	66.619
40	20.707	22.164	24.433	26.509	29.051	32.345	34.872	37.134	39.335	41.622	44.165	47.269	51.805	55.758	59.342	63.691	66.766	73.403
50	27.991	29.707	32.357	34.764	37.689	41.449	44.313	46.864	49.335	51.892	54.723	58.164	63.167	67.505	71.420	76.154	79.490	86.660
60	35.534	37.485	40.482	43.188	46.459	50.641	53.809	56.620	59.335	62.135	65.226	68.972	74.397	79.082	83.298	88.379	91.952	99.608
70	43.275	45.442	48.758	51.739	55.329	59.898	63.346	66.396	69.334	72.358	75.689	79.715	85.527	90.531	95.023	100.425	104.215	112.317
80	51.172	53.540	57.153	60.391	64.278	69.207	72.915	76.188	79.334	82.566	86.120	90.405	96.578	101.879	106.629	112.329	116.321	124.839
90	59.196	61.754	65.647	69.126	73.291	78.558	82.511	85.993	89.334	92.761	96.524	101.054	107.565	113.145	118.136	124.116	128.299	137.208
100	67.328	70.065	74.222	77.929	82.358	87.945	92.129	95.808	99.334	102.946	106.906	111.667	118.498	124.342	129.561	135.807	140.170	149.449

Table B.5 Critical values for the F distribution

The table contains $F_{\nu_1, \nu_2; \alpha}$ values for given values of ν_1, ν_2 and α with

$$P(F_{\nu_1, \nu_2} \geq F_{\nu_1, \nu_2; \alpha}) = \alpha.$$

For the values of α for which $F_{\nu_1, \nu_2; \alpha}$ is not tabulated, but $F_{\nu_1, \nu_2; 1-\alpha}$ is, the relationship used is $F_{\nu_1, \nu_2; \alpha} = \dfrac{1}{F_{\nu_2, \nu_1; 1-\alpha}}$.



$\alpha = 0.10$

ν_2	ν_1																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.11
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

Table B.5 Critical values for the F distribution (continued)

$\alpha = 0.05$

ν_2	ν_1																	
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Table B.5 Critical values for the F distribution (continued)

$\alpha = 0.025$

ν_2	ν_1																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.41	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

Table B.5 Critical values for the F distribution (continued)

$\alpha = 0.01$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99 48	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3 51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4 13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2 41	2.32	2.18	2.04	1.88	1.79	1.70	1 59	1.47	1.32	1.00

Table B.5 Critical values for the F distribution (continued)

$\alpha = 0.005$

ν_2	ν_1																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	16211	20000	21615	22500	23056	23437	23715	23925	24091	24224	24426	24630	24836	24940	25044	25148	25253	25359	25465
2	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4	199.4	199.4	199.4	199.5	199.5	199.5	199.5	199.5	199.5	199.5
3	55.55	49.80	47.47	46.20	45.39	44.84	44.43	44.13	43.88	43.68	43.39	43.08	42.78	42.62	42.47	42.31	42.15	41.99	41.83
4	31.33	26.28	24.26	23.15	22.46	21.98	21.62	21.35	21.14	20.97	20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	19.32
5	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	13.38	13.15	12.90	12.78	12.66	12.53	12.40	12.27	12.14
6	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	8.88
7	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	8.18	7.97	7.75	7.64	7.53	7.42	7.31	7.19	7.08
8	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	5.95
9	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30	5.19
10	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.66	5.47	5.27	5.17	5.07	4.97	4.86	4.75	4.64
11	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	5.24	5.05	4.86	4.76	4.65	4.55	4.45	4.34	4.23
12	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.91	4.72	4.53	4.43	4.33	4.23	4.12	4.01	3.90
13	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.64	4.46	4.27	4.17	4.07	3.97	3.87	3.76	3.65
14	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.43	4.25	4.06	3.96	3.86	3.76	3.66	3.55	3.44
15	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.25	4.07	3.88	3.79	3.69	3.59	3.48	3.37	3.26
16	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	4.10	3.92	3.73	3.64	3.54	3.44	3.33	3.22	3.11
17	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	3.97	3.79	3.61	3.51	3.41	3.31	3.21	3.10	2.98
18	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	3.86	3.68	3.50	3.40	3.30	3.20	3.10	2.99	2.87
19	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	3.76	3.59	3.40	3.31	3.21	3.11	3.00	2.89	2.78
20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	2.69
21	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77	3.60	3.43	3.24	3.15	3.05	2.95	2.84	2.73	2.61
22	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70	3.54	3.36	3.18	3.08	2.98	2.88	2.77	2.66	2.55
23	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64	3.47	3.30	3.12	3.02	2.92	2.82	2.71	2.60	2.48
24	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55	2.43
25	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54	3.37	3.20	3.01	2.92	2.82	2.72	2.61	2.50	2.38
26	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49	3.33	3.15	2.97	2.87	2.77	2.67	2.56	2.45	2.33
27	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45	3.28	3.11	2.93	2.83	2.73	2.63	2.52	2.41	2.29
28	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41	3.25	3.07	2.89	2.79	2.69	2.59	2.48	2.37	2.25
29	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48	3.38	3.21	3.04	2.86	2.76	2.66	2.56	2.45	2.33	2.21
30	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.18	3.01	2.82	2.73	2.63	2.52	2.42	2.30	2.18
40	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	2.95	2.78	2.60	2.50	2.40	2.30	2.18	2.06	1.93
60	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.74	2.57	2.39	2.29	2.19	2.08	1.96	1.83	1.69
120	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.54	2.37	2.19	2.09	1.98	1.87	1.75	1.61	1.43
∞	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.36	2.19	2.00	1.90	1.79	1.67	1.53	1.36	1.00

Table B.6 Cumulative binomial distibution

The table contains the values for $F(b)$ of the bi(n,π) distribution and for the given values of n and π with

$$F(b) = P(X \leq b).$$

For $\pi > 0.5$ the relationship used is

$$F(b) = 1 - G(n - b - 1)$$

with F the cumulative distribution function of the bi(n, π) distribution and G the cumulative distribution function of the bi$(n, 1-\pi)$ distribution.

		π						
n	b	0.05	0.10	0.20	0.25	0.30	0.40	0.50
5	0	0.77378	0.59049	0.32768	0.23730	0.16807	0.07776	0.03125
	1	0.97741	0.91854	0.73728	0.63281	0.52822	0.33696	0.18750
	2	0.99884	0.99144	0.94208	0.89648	0.83692	0.68256	0.50000
	3	0.99997	0.99954	0.99328	0.98438	0.96922	0.91296	0.81250
	4	1.00000	0.99999	0.99968	0.99902	0.99757	0.98976	0.96875
	5		1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
8	0	0.66342	0.43047	0.16777	0.10011	0.05765	0.01680	0.00391
	1	0.94276	0.81310	0.50332	0.36708	0.25530	0.10638	0.03516
	2	0.99421	0.96191	0.79692	0.67854	0.55177	0.31539	0.14453
	3	0.99963	0.99498	0.94372	0.88618	0.80590	0.59409	0.36328
	4	0.99998	0.99957	0.98959	0.97270	0.94203	0.82633	0.63672
	5	1.00000	0.99998	0.99877	0.99577	0.98871	0.95019	0.85547
	6		1.00000	0.99992	0.99962	0.99871	0.99148	0.96484
	7			1.00000	0.99998	0.99993	0.99934	0.99609
	8				1.00000	1.00000	1.00000	1.00000
10	0	0.59874	0.34868	0.10737	0.05631	0.02825	0.00605	0.00098
	1	0.91386	0.73610	0.37581	0.24403	0.14931	0.04636	0.01074
	2	0.98850	0.92981	0.67780	0.52559	0.38278	0.16729	0.05469
	3	0.99897	0.98720	0.87913	0.77588	0.64961	0.38228	0.17188
	4	0.99994	0.99837	0.96721	0.92187	0.84973	0.63310	0.37695
	5	1.00000	0.99985	0.99363	0.98027	0.95265	0.83376	0.62305
	6		0.99999	0.99914	0.99649	0.98941	0.94524	0.82813
	7		1.00000	0.99992	0.99958	0.99841	0.98771	0.94531
	8			1.00000	0.99997	0.99986	0.99832	0.98926
	9				1.00000	0.99999	0.99990	0.99902
	10					1.00000	1.00000	1.00000
12	0	0.54036	0.28243	0.06872	0.03168	0.01384	0.00218	0.00024
	1	0.88164	0.65900	0.27488	0.15838	0.08503	0.01959	0.00317
	2	0.98043	0.88913	0.55835	0.39068	0.25282	0.08344	0.01929
	3	0.99776	0.97436	0.79457	0.64878	0.49252	0.22534	0.07300
	4	0.99982	0.99567	0.92744	0.84236	0.72366	0.43818	0.19385
	5	0.99999	0.99946	0.98059	0.94560	0.88215	0.66521	0.38721
	6	1.00000	0.99995	0.99610	0.98575	0.96140	0.84179	0.61279
	7		1.00000	0.99942	0.99722	0.99051	0.94269	0.80615
	8			0.99994	0.99961	0.99831	0.98473	0.92700
	9			1.00000	0.99996	0.99979	0.99719	0.98071
	10				1.00000	0.99998	0.99968	0.99683
	11					1.00000	0.99998	0.99976
	12						1.00000	1.00000

Table B.6 Cumulative binomial distribution (continued)

		π						
n	b	0.05	0.10	0.20	0.25	0.30	0.40	0.50
15	0	0.46329	0.20589	0.03518	0.01336	0.00475	0.00047	0.00003
	1	0.82905	0.54904	0.16713	0.08018	0.03527	0.00517	0.00049
	2	0.96380	0.81594	0.39802	0.23609	0.12683	0.02711	0.00369
	3	0.99453	0.94444	0.64816	0.46129	0.29687	0.09050	0.01758
	4	0.99939	0.98728	0.83577	0.68649	0.51549	0.21728	0.05923
	5	0.99995	0.99775	0.93895	0.85163	0.72162	0.40322	0.15088
	6	1.00000	0.99969	0.98194	0.94338	0.86886	0.60981	0.30362
	7		0.99997	0.99576	0.98270	0.94999	0.78690	0.50000
	8		1.00000	0.99922	0.99581	0.98476	0.90495	0.69638
	9			0.99989	0.99921	0.99635	0.96617	0.84912
	10			0.99999	0.99988	0.99933	0.99065	0.94077
	11			1.00000	0.99999	0.99991	0.99807	0.98242
	12				1.00000	0.99999	0.99972	0.99631
	13					1.00000	0.99997	0.99951
	14						1.00000	0.99997
	15							1.00000
18	0	0.39721	0.15009	0.01801	0.00564	0.00163	0.00010	0.00000
	1	0.77352	0.45028	0.09908	0.03946	0.01419	0.00132	0.00007
	2	0.94187	0.73380	0.27134	0.13531	0.05995	0.00823	0.00066
	3	0.98913	0.90180	0.50103	0.30569	0.16455	0.03278	0.00377
	4	0.99845	0.97181	0.71635	0.51867	0.33265	0.09417	0.01544
	5	0.99983	0.99358	0.86708	0.71745	0.53438	0.20876	0.04813
	6	0.99998	0.99883	0.94873	0.86102	0.72170	0.37428	0.11894
	7	1.00000	0.99983	0.98372	0.94305	0.85932	0.56344	0.24034
	8		0.99998	0.99575	0.98065	0.94041	0.73684	0.40726
	9		1.00000	0.99909	0.99458	0.97903	0.86529	0.59274
	10			0.99984	0.99876	0.99393	0.94235	0.75966
	11			0.99998	0.99977	0.99857	0.97972	0.88106
	12			1.00000	0.99997	0.99973	0.99425	0.95187
	13				1.00000	0.99996	0.99872	0.98456
	14					1.00000	0.99979	0.99623
	15						0.99997	0.99934
	16						1.00000	0.99993
	17							1.00000
	18							
20	0	0.35849	0.12158	0.01153	0.00317	0.00080	0.00004	0.00000
	1	0.73584	0.39175	0.06918	0.02431	0.00764	0.00052	0.00002
	2	0.92452	0.67693	0.20608	0.09126	0.03548	0.00361	0.00020
	3	0.98410	0.86705	0.41145	0.22516	0.10709	0.01596	0.00129
	4	0.99743	0.95683	0.62965	0.41484	0.23751	0.05095	0.00591
	5	0.99967	0.98875	0.80421	0.61717	0.41637	0.12560	0.02069
	6	0.99997	0.99761	0.91331	0.78578	0.60801	0.25001	0.05766
	7	1.00000	0.99958	0.96786	0.89819	0.77227	0.41589	0.13159
	8		0.99994	0.99002	0.95907	0.88667	0.59560	0.25172
	9		0.99999	0.99741	0.98614	0.95204	0.75534	0.41190
	10		1.00000	0.99944	0.99606	0.98286	0.87248	0.58810
	11			0.99990	0.99906	0.99486	0.94347	0.74828
	12			0.99998	0.99982	0.99872	0.97897	0.86841
	13			1.00000	0.99997	0.99974	0.99353	0.94234
	14				1.00000	0.99996	0.99839	0.97931
	15					0.99999	0.99968	0.99409
	16					1.00000	0.99995	0.99871
	17						0.99999	0.99980
	18						1.00000	0.99998
	19							1.00000
	20							

Table B.6 Cumulative binomial distribution (continued)

n	b	π						
		0.05	0.10	0.20	0.25	0.30	0.40	0.50
25	0	0.27739	0.07179	0.00378	0.00075	0.00013	0.00000	0.00000
	1	0.64238	0.27121	0.02739	0.00702	0.00157	0.00005	0.00000
	2	0.87289	0.53709	0.09823	0.03211	0.00896	0.00043	0.00001
	3	0.96591	0.76359	0.23399	0.09621	0.03324	0.00237	0.00008
	4	0.99284	0.90201	0.42067	0.21374	0.09047	0.00947	0.00046
	5	0.99879	0.96660	0.61669	0.37828	0.19349	0.02936	0.00204
	6	0.99983	0.99052	0.78004	0.56110	0.34065	0.07357	0.00732
	7	0.99998	0.99774	0.89088	0.72651	0.51185	0.15355	0.02164
	8	1.00000	0.99954	0.95323	0.85056	0.67693	0.27353	0.05388
	9		0.99992	0.98267	0.92867	0.81056	0.42462	0.11476
	10		0.99999	0.99445	0.97033	0.90220	0.58577	0.21218
	11		1.00000	0.99846	0.98927	0.95575	0.73228	0.34502
	12			0.99963	0.99663	0.98253	0.84623	0.50000
	13			0.99992	0.99908	0.99401	0.92220	0.65498
	14			0.99999	0.99979	0.99822	0.96561	0.78782
	15			1.00000	0.99996	0.99955	0.98683	0.88524
	16				0.99999	0.99990	0.99567	0.94612
	17				1.00000	0.99998	0.99879	0.97836
	18					1.00000	0.99972	0.99268
	19						0.99995	0.99796
	20						0.99999	0.99954
	21						1.00000	0.99992
	22							0.99999
	23							1.00000
	24							
	25							
30	0	0.21464	0.04239	0.00124	0.00018	0.00002	0.00000	0.00000
	1	0.55354	0.18370	0.01052	0.00196	0.00031	0.00000	0.00000
	2	0.81218	0.41135	0.04418	0.01060	0.00211	0.00005	0.00000
	3	0.93923	0.64744	0.12271	0.03745	0.00932	0.00031	0.00000
	4	0.98436	0.82451	0.25523	0.09787	0.03015	0.00151	0.00003
	5	0.99672	0.92681	0.42751	0.20260	0.07659	0.00566	0.00016
	6	0.99943	0.97417	0.60697	0.34805	0.15952	0.01718	0.00072
	7	0.99992	0.99222	0.76079	0.51429	0.28138	0.04352	0.00261
	8	0.99999	0.99798	0.87135	0.67360	0.43152	0.09401	0.00806
	9	1.00000	0.99955	0.93891	0.80341	0.58881	0.17629	0.02139
	10		0.99991	0.97438	0.89427	0.73037	0.29147	0.04937
	11		0.99998	0.99051	0.94934	0.84068	0.43109	0.10024
	12		1.00000	0.99689	0.97841	0.91553	0.57847	0.18080
	13			0.99910	0.99182	0.95995	0.71450	0.29233
	14			0.99977	0.99725	0.98306	0.82463	0.42777
	15			0.99995	0.99918	0.99363	0.90294	0.57223
	16			0.99999	0.99978	0.99788	0.95189	0.70767
	17			1.00000	0.99995	0.99937	0.97876	0.81920
	18				0.99999	0.99984	0.99170	0.89976
	19				1.00000	0.99996	0.99715	0.95063
	20					0.99999	0.99914	0.97861
	21					1.00000	0.99978	0.99194
	22						0.99995	0.99739
	23						0.99999	0.99928
	24						1.00000	0.99984
	25							0.99997
	26							1.00000
	27							
	28							
	29							
	30							

Table B.7 Cumulative Poisson distribution

The table contains the values for $F(b)$ for the po(λ) distribution for given values of λ and b with $F(b) = P(X \leq b)$.

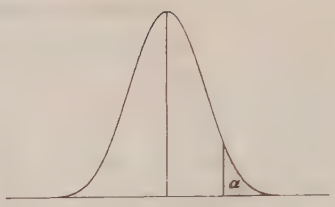
	λ						
b	0.1	0.2	0.3	0.4	0.5	0.6	0.7
0	0.90484	0.81873	0.74082	0.67032	0.60653	0.54881	0.49659
1	0.99532	0.98248	0.96306	0.93845	0.90980	0.87810	0.84420
2	0.99985	0.99885	0.99640	0.99207	0.98561	0.97688	0.96586
3	1.00000	0.99994	0.99973	0.99922	0.99825	0.99664	0.99425
4		1.00000	0.99998	0.99994	0.99983	0.99961	0.99921
5			1.00000	1.00000	0.99999	0.99996	0.99991
6					1.00000	1.00000	0.99999
7							1.00000

	λ						
b	0.8	0.9	1.0	1.5	2.0	2.5	3.0
0	0.44933	0.40657	0.36788	0.22313	0.13534	0.08208	0.04979
1	0.80879	0.77248	0.73576	0.55783	0.40601	0.28730	0.19915
2	0.95258	0.93714	0.91970	0.80885	0.67668	0.54381	0.42319
3	0.99092	0.98654	0.98101	0.93436	0.85712	0.75758	0.64723
4	0.99859	0.99766	0.99634	0.98142	0.94735	0.89118	0.81526
5	0.99982	0.99966	0.99941	0.99554	0.98344	0.95798	0.91608
6	0.99998	0.99996	0.99992	0.99907	0.99547	0.98581	0.96649
7	1.00000	1.00000	0.99999	0.99983	0.99890	0.99575	0.98810
8			1.00000	0.99997	0.99976	0.99886	0.99620
9				1.00000	0.99995	0.99972	0.99890
10					0.99999	0.99994	0.99971
11					1.00000	0.99999	0.99993
12						1.00000	0.99998
13							1.00000

	λ						
b	4	5	6	7	8	9	10
0	0.01832	0.00674	0.00248	0.00091	0.00034	0.00012	0.00005
1	0.09158	0.04043	0.01735	0.00730	0.00302	0.00123	0.00050
2	0.23810	0.12465	0.06197	0.02964	0.01375	0.00623	0.00277
3	0.43347	0.26503	0.15120	0.08177	0.04238	0.02123	0.01034
4	0.62884	0.44049	0.28506	0.17299	0.09963	0.05496	0.02925
5	0.78513	0.61596	0.44568	0.30071	0.19124	0.11569	0.06709
6	0.88933	0.76218	0.60630	0.44971	0.31337	0.20678	0.13014
7	0.94887	0.86663	0.74398	0.59871	0.45296	0.32390	0.22022
8	0.97864	0.93191	0.84724	0.72909	0.59255	0.45565	0.33282
9	0.99187	0.96817	0.91608	0.83050	0.71662	0.58741	0.45793
10	0.99716	0.98630	0.95738	0.90148	0.81589	0.70599	0.58304
11	0.99908	0.99455	0.97991	0.94665	0.88808	0.80301	0.69678
12	0.99973	0.99798	0.99117	0.97300	0.93620	0.87577	0.79156
13	0.99992	0.99930	0.99637	0.98719	0.96582	0.92615	0.86446
14	0.99998	0.99977	0.99860	0.99428	0.98274	0.95853	0.91654
15	1.00000	0.99993	0.99949	0.99759	0.99177	0.97796	0.95126
16		0.99998	0.99983	0.99904	0.99628	0.98889	0.97296
17		0.99999	0.99994	0.99964	0.99841	0.99468	0.98572
18		1.00000	0.99998	0.99987	0.99935	0.99757	0.99281
19			0.99999	0.99996	0.99975	0.99894	0.99655
20			1.00000	0.99999	0.99991	0.99956	0.99841
21				1.00000	0.99997	0.99983	0.99930
22					0.99999	0.99993	0.99970
23					1.00000	0.99998	0.99988
24						1.00000	0.99995
25							1.00000

Table B.8 Critical values for the $r(n)$ distibution

The table contains the values of $r_{n,\alpha}$ for given n and α values with $P(R_n \geq r_{n,\alpha}) = \alpha$.



n	α					
	0.25	0.10	0.05	0.025	0.01	0.005
3	0.70711	0.95106	0.98769	0.99692	0.99951	0.99988
4	0.50000	0.80000	0.90000	0.95000	0.98000	0.99000
5	0.40397	0.68705	0.80538	0.87834	0.93433	0.95873
6	0.34730	0.60840	0.72930	0.81140	0.88219	0.91720
7	0.30907	0.55086	0.66944	0.75449	0.83287	0.87453
8	0.28113	0.50673	0.62149	0.70673	0.78872	0.83434
9	0.25957	0.47159	0.58221	0.66638	0.74978	0.79768
10	0.24230	0.44280	0.54936	0.63190	0.71546	0.76459
11	0.22807	0.41866	0.52140	0.60207	0.68509	0.73479
12	0.21607	0.39806	0.49726	0.57598	0.65807	0.70789
13	0.20579	0.38022	0.47616	0.55294	0.63386	0.68353
14	0.19684	0.36456	0.45750	0.53241	0.61205	0.66138
15	0.18897	0.35069	0.44086	0.51398	0.59227	0.64115
16	0.18197	0.33828	0.42590	0.49731	0.57425	0.62259
17	0.17569	0.32710	0.41236	0.48215	0.55774	0.60551
18	0.17002	0.31696	0.40003	0.46828	0.54255	0.58972
19	0.16487	0.30770	0.38873	0.45553	0.52852	0.57507
20	0.16015	0.29921	0.37834	0.44376	0.51550	0.56144
21	0.15582	0.29138	0.36874	0.43286	0.50340	0.54871
22	0.15183	0.28414	0.35983	0.42271	0.49209	0.53680
23	0.14812	0.27741	0.35153	0.41325	0.48151	0.52562
24	0.14468	0.27114	0.34378	0.40439	0.47158	0.51510
25	0.14146	0.26527	0.33652	0.39607	0.46223	0.50518
26	0.13845	0.25977	0.32970	0.38824	0.45341	0.49581
27	0.13562	0.25459	0.32328	0.38086	0.44508	0.48693
28	0.13296	0.24972	0.31722	0.37389	0.43718	0.47851
29	0.13045	0.24511	0.31149	0.36728	0.42969	0.47051
30	0.12808	0.24075	0.30606	0.36101	0.42257	0.46289
31	0.12583	0.23661	0.30090	0.35505	0.41579	0.45563
32	0.12370	0.23268	0.29599	0.34937	0.40933	0.44870
33	0.12167	0.22894	0.29132	0.34396	0.40315	0.44207
34	0.11974	0.22537	0.28686	0.33879	0.39725	0.43573
35	0.11789	0.22197	0.28259	0.33384	0.39160	0.42965
40	0.10980	0.20699	0.26381	0.31201	0.36655	0.40264
45	0.10318	0.19469	0.24833	0.29396	0.34575	0.38014
50	0.09763	0.18434	0.23529	0.27871	0.32813	0.36103
55	0.09288	0.17549	0.22411	0.26561	0.31295	0.34453
60	0.08877	0.16780	0.21438	0.25420	0.29970	0.33010
65	0.08516	0.16104	0.20582	0.24415	0.28799	0.31734
70	0.08196	0.15504	0.19821	0.23520	0.27756	0.30596
75	0.07909	0.14966	0.19139	0.22716	0.26819	0.29571
80	0.07650	0.14480	0.18522	0.21990	0.25970	0.28643
85	0.07416	0.14039	0.17961	0.21329	0.25197	0.27797
90	0.07201	0.13636	0.17449	0.20725	0.24490	0.27022
95	0.07004	0.13266	0.16978	0.20169	0.23838	0.26308
100	0.06823	0.12924	0.16543	0.19655	0.23236	0.25648

Table B.9 The Z-transformation for the correlation coefficient

The table contains the transformed values $z(r)$ for the given values of r where

$$z(r) = \frac{1}{2} ln \left(\frac{1+r}{1-r} \right).$$

For negative values of r the relationship used is

$$z(-r) = -z(r).$$

r	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.01000	0.02000	0.03001	0.04002	0.05004	0.06007	0.07011	0.08017	0.09024
0.1	0.10034	0.11045	0.12058	0.13074	0.14093	0.15114	0.16139	0.17167	0.18198	0.19234
0.2	0.20273	0.21317	0.22366	0.23419	0.24477	0.25541	0.26611	0.27686	0.28768	0.29857
0.3	0.30952	0.32055	0.33165	0.34283	0.35409	0.36544	0.37689	0.38842	0.40006	0.41180
0.4	0.42365	0.43561	0.44769	0.45990	0.47223	0.48470	0.49731	0.51007	0.52298	0.53606
0.5	0.54931	0.56273	0.57634	0.59015	0.60416	0.61838	0.63283	0.64752	0.66246	0.67767
0.6	0.69315	0.70892	0.72501	0.74142	0.75817	0.77530	0.79281	0.81074	0.82911	0.84796
0.7	0.86730	0.88718	0.90764	0.92873	0.95048	0.97296	0.99622	1.02033	1.04537	1.07143
0.8	1.09861	1.12703	1.15682	1.18814	1.22117	1.25615	1.29334	1.33308	1.37577	1.42193
0.9	1.47222	1.52752	1.58903	1.65839	1.73805	1.83178	1.94591	2.09230	2.29756	2.64665

r	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.90	1.47222	1.47751	1.48285	1.48824	1.49368	1.49918	1.50473	1.51034	1.51601	1.52174
0.91	1.52752	1.53337	1.53928	1.54526	1.55130	1.55741	1.56359	1.56984	1.57616	1.58256
0.92	1.58903	1.59558	1.60221	1.60892	1.61571	1.62260	1.62957	1.63663	1.64379	1.65104
0.93	1.65839	1.66584	1.67340	1.68107	1.68885	1.69674	1.70475	1.71288	1.72114	1.72953
0.94	1.73805	1.74671	1.75552	1.76447	1.77358	1.78284	1.79227	1.80188	1.81166	1.82162
0.95	1.83178	1.84214	1.85270	1.86349	1.87450	1.88574	1.89723	1.90898	1.92100	1.93331
0.96	1.94591	1.95882	1.97207	1.98566	1.99961	2.01395	2.02870	2.04388	2.05952	2.07565
0.97	2.09230	2.10950	2.12730	2.14574	2.16486	2.18472	2.20539	2.22692	2.24940	2.27291
0.98	2.29756	2.32346	2.35074	2.37958	2.41014	2.44266	2.47741	2.51472	2.55499	2.59875
0.99	2.64665	2.69958	2.75873	2.82574	2.90307	2.99448	3.10630	3.25039	3.45338	3.80020

Table B.10 Critical values for the sign test

The table contains the values for $b_{n,\alpha}$ (lower critical values) for the given values of n and α.

Here $b_{n,\alpha}$ is the largest integer number for which

$$P(B \le b_{n,\alpha}) \le \alpha.$$

The upper critical values are given by: $n - b_{n,\alpha}$

	α					
n	0.25	0.10	0.05	0.025	0.01	0.005
3	0					
4	0	0				
5	1	0	0			
6	1	0	0	0		
7	2	1	0	0	0	
8	2	1	1	0	0	0
9	2	2	1	1	0	0
10	3	2	1	1	0	0
11	3	2	2	1	1	0
12	4	3	2	2	1	1
13	4	3	3	2	1	1
14	5	4	3	2	2	1
15	5	4	3	3	2	2
16	6	4	4	3	2	2
17	6	5	4	4	3	2
18	7	5	5	4	3	3
19	7	6	5	4	4	3
20	7	6	5	5	4	3
21	8	7	6	5	4	4
22	8	7	6	5	5	4
23	9	7	7	6	5	4
24	9	8	7	6	5	5
25	10	8	7	7	6	5
26	10	9	8	7	6	6
27	11	9	8	7	7	6
28	11	10	9	8	7	6
29	12	10	9	8	7	7
30	12	10	10	9	8	7
31	13	11	10	9	8	7
32	13	11	10	9	8	8
33	14	12	11	10	9	8
34	14	12	11	10	9	9
35	15	13	12	11	10	9
36	15	13	12	11	10	9
37	15	14	13	12	10	10
38	16	14	13	12	11	10
39	16	15	13	12	11	11
40	17	15	14	13	12	11

Table B.11 Critical values of the signed rank test

The table contains the values of $s_{n,\alpha}$ (lower critical values) for the given values of n and α.

Here $s_{n,\alpha}$ is the largest integer for which

$$P(S \le s_{n,\alpha}) \le \alpha.$$

The upper critical values are given by: $\frac{n(n+1)}{2} - s_{n,\alpha}$

	α					
n	0.25	0.1	0.05	0.025	0.01	0.005
3	1					
4	2	0				
5	4	2	0			
6	6	3	2	0		
7	9	5	3	2	0	
8	12	8	5	3	1	0
9	16	10	8	5	3	1
10	20	14	10	8	5	3
11	24	17	13	10	7	5
12	29	21	17	13	9	7
13	35	26	21	17	12	9
14	40	31	25	21	15	12
15	47	36	30	25	19	15
16	54	42	35	29	23	19
17	61	48	41	34	27	23
18	69	55	47	40	32	27
19	77	62	53	46	37	32
20	86	69	60	52	43	37

Table B.12 Critical values for the rank sum test

The table contains the values of $w_{n_1, n_2; \alpha}$ (lower critical values) for the given values of n_1, n_2 and α with $n_1 \le n_2$. Here $w_{n_1, n_2; \alpha}$ is the largest integer for which

$$P(W \le w_{n_1, n_2; \alpha}) \le \alpha.$$

The upper critical values are given by: $n_1 n_2 - w_{n_1, n_2; \alpha}$

		α					
n_2	n_1	0.25	0.10	0.05	0.025	0.01	0.005
3	1	0					
	2	1	0				
	3	2	1	0			
4	1	0					
	2	1	0				
	3	3	1	0			
	4	5	3	1	0		
5	1	0					
	2	2	1	0			
	3	4	2	1	0		
	4	6	4	2	1	0	
	5	8	5	4	2	1	0
6	1	0					
	2	3	1	0			
	3	5	3	2	1		
	4	8	5	3	2	1	0
	5	10	7	5	3	2	1
	6	13	9	7	5	3	2
7	1	1					
	2	4	1	0			
	3	6	4	2	1	0	
	4	9	6	4	3	1	0
	5	12	8	6	5	3	1
	6	15	11	8	6	4	3
	7	18	13	11	8	6	4

		α					
n_2	n_1	0.25	0.1	0.05	0.025	0.01	0.005
8	1	1					
	2	4	2	1	0		
	3	8	5	3	2	0	
	4	11	7	5	4	2	1
	5	14	10	8	6	4	2
	6	18	13	10	8	6	4
	7	21	16	13	10	7	6
	8	24	19	15	13	9	7
9	1	1	0				
	2	5	2	1	0		
	3	9	5	4	2	1	0
	4	12	9	6	4	3	1
	5	16	12	9	7	5	3
	6	20	15	12	10	7	5
	7	24	18	15	12	9	7
	8	28	22	18	15	11	9
	9	32	25	21	17	14	11
10	1	1	0				
	2	6	3	1	0		
	3	10	6	4	3	1	0
	4	14	10	7	5	3	2
	5	18	13	11	8	6	4
	6	23	17	14	11	8	6
	7	27	21	17	14	11	9
	8	31	24	20	17	13	11
	9	36	28	24	20	16	13
	10	40	32	27	23	19	16

Table B.13 a **values for** $(I \times J)$ **log linear analysis**

In this table the a values used in

$$s^2_l = \begin{cases} \sum_{r=1}^{I} \sum_{s=1}^{J} \dfrac{a_{rs}^2}{I^2 J^2 w_{ij}} & \text{(saturated model)} \\ \sum_{r=1}^{I} \sum_{s=1}^{J} \dfrac{a_{rs}^2}{I^2 J^2 v_{ij}} & \text{(independence model)} \end{cases}$$

are tabulated for the different estimated effects namely for the values of $I \le J$ with $I \ge 2$, $I + J \le 8$.

(2 × 2) table

		a_{11}	a_{12}	a_{21}	a_{22}
l^A	1	1	1	−1	−1
	2	−1	−1	1	1
l^B	1	1	−1	1	−1
	2	−1	1	−1	1
l^{AB}	11	1	−1	−1	1
	12	−1	1	1	−1
	21	−1	1	1	−1
	22	1	−1	−1	1

(3 × 3) table

		a_{11}	a_{12}	a_{13}	a_{21}	a_{22}	a_{23}	a_{31}	a_{32}	a_{33}
l^A	1	2	2	2	−1	−1	−1	−1	−1	−1
	2	−1	−1	−1	2	2	2	−1	−1	−1
	3	−1	−1	−1	−1	−1	−1	2	2	2
l^B	1	2	−1	−1	2	−1	−1	2	−1	−1
	2	−1	2	−1	−1	2	−1	−1	2	−1
	3	−1	−1	2	−1	−1	2	−1	−1	2
l^{AB}	11	4	−2	−2	−2	1	1	−2	1	1
	12	−2	4	−2	1	−2	1	1	−2	1
	13	−2	−2	4	1	1	−2	1	1	−2
	21	−2	1	1	4	−2	−2	−2	1	1
	22	1	−2	1	−2	4	−2	1	−2	1
	23	1	1	−2	−2	−2	4	1	1	−2
	31	−2	1	1	−2	1	1	4	−2	−2
	32	1	−2	1	1	−2	1	−2	4	−2
	33	1	1	−2	1	1	−2	−2	−2	4

(2 × 3) table

		a_{11}	a_{12}	a_{13}	a_{21}	a_{22}	a_{23}
l^A	1	1	1	1	−1	−1	−1
	2	−1	−1	−1	1	1	1
l^B	1	2	−1	−1	2	−1	−1
	2	−1	2	−1	−1	2	−1
	3	−1	−1	2	−1	−1	2
l^{AB}	11	2	−1	−1	−2	1	1
	12	−1	2	−1	1	−2	1
	13	−1	−1	2	1	1	−2
	21	−2	1	1	2	−1	−1
	22	1	−2	1	−1	2	−1
	23	1	1	−2	−1	−1	2

(2 × 4) table

		a_{11}	a_{12}	a_{13}	a_{14}	a_{21}	a_{22}	a_{23}	a_{24}
l^A	1	1	1	1	1	−1	−1	−1	−1
	2	−1	−1	−1	−1	1	1	1	1
l^B	1	3	−1	−1	−1	3	−1	−1	−1
	2	−1	3	−1	−1	−1	3	−1	−1
	3	−1	−1	3	−1	−1	−1	3	−1
	4	−1	−1	−1	3	−1	−1	−1	3
l^{AB}	11	3	−1	−1	−1	−3	1	1	1
	12	−1	3	−1	−1	1	−3	1	1
	13	−1	−1	3	−1	1	1	−3	1
	14	−1	−1	−1	3	1	1	1	−3
	21	−3	1	1	1	3	−1	−1	−1
	22	1	−3	1	1	−1	3	−1	−1
	23	1	1	−3	1	−1	−1	3	−1
	24	1	1	1	−3	−1	−1	−1	3

Table B.13 a **values for** $(I \times J)$ **log linear analysis** (continued)

(3 × 4) table

		a_{11}	a_{12}	a_{13}	a_{14}	a_{21}	a_{22}	a_{23}	a_{24}	a_{31}	a_{32}	a_{33}	a_{34}
l^A	1	2	2	2	2	−1	−1	−1	−1	−1	−1	−1	−1
	2	−1	−1	−1	−1	2	2	2	2	−1	−1	−1	−1
	3	−1	−1	−1	−1	−1	−1	−1	−1	2	2	2	2
l^B	1	3	−1	−1	−1	3	−1	−1	−1	3	−1	−1	−1
	2	−1	3	−1	−1	−1	3	−1	−1	−1	3	−1	−1
	3	−1	−1	3	−1	−1	−1	3	−1	−1	−1	3	−1
	4	−1	−1	−1	3	−1	−1	−1	3	−1	−1	−1	3
l^{AB}	11	6	−2	−2	−2	−3	1	1	1	−3	1	1	1
	12	−2	6	−2	−2	1	−3	1	1	1	−3	1	1
	13	−2	−2	6	−2	1	1	−3	1	1	1	−3	1
	14	−2	−2	−2	6	1	1	1	−3	1	1	1	−3
	21	−3	1	1	1	6	−2	−2	−2	−3	1	1	1
	22	1	−3	1	1	−2	6	−2	−2	1	−3	1	1
	23	1	1	−3	1	−2	−2	6	−2	1	1	−3	1
	24	1	1	1	−3	−2	−2	−2	6	1	1	1	−3
	31	−3	1	1	1	−3	1	1	1	6	−2	−2	−2
	32	1	−3	1	1	1	−3	1	1	−2	6	−2	−2
	33	1	1	−3	1	1	1	−3	1	−2	−2	6	−2
	34	1	1	1	−3	1	1	1	−3	−2	−2	−2	6

(2 × 5) table

		a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{21}	a_{22}	a_{23}	a_{24}	a_{25}
l^A	1	1	1	1	1	1	−1	−1	−1	−1	−1
	2	−1	−1	−1	−1	−1	1	1	1	1	1
l^B	1	4	−1	−1	−1	−1	4	−1	−1	−1	−1
	2	−1	4	−1	−1	−1	−1	4	−1	−1	−1
	3	−1	−1	4	−1	−1	−1	−1	4	−1	−1
	4	−1	−1	−1	4	−1	−1	−1	−1	4	−1
	5	−1	−1	−1	−1	4	−1	−1	−1	−1	4
l^{AB}	11	4	−1	−1	−1	−1	−4	1	1	1	1
	12	−1	4	−1	−1	−1	1	−4	1	1	1
	13	−1	−1	4	−1	−1	1	1	−4	1	1
	14	−1	−1	−1	4	−1	1	1	1	−4	1
	15	−1	−1	−1	−1	4	1	1	1	1	−4
	21	−4	1	1	1	1	4	−1	−1	−1	−1
	22	1	−4	1	1	1	−1	4	−1	−1	−1
	23	1	1	−4	1	1	−1	−1	4	−1	−1
	24	1	1	1	−4	1	−1	−1	−1	4	−1
	25	1	1	1	1	−4	−1	−1	−1	−1	4

Table B.13 a **values for** $(I \times J)$ **log linear analysis** (continued)

(2 × 6) table

		a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}
l^A	1	1	1	1	1	1	1	–1	–1	–1	–1	–1	–1
	2	–1	–1	–1	–1	–1	–1	1	1	1	1	1	1
l^B	1	5	–1	–1	–1	–1	–1	5	–1	–1	–1	–1	–1
	2	–1	5	–1	–1	–1	–1	–1	5	–1	–1	–1	–1
	3	–1	–1	5	–1	–1	–1	–1	–1	5	–1	–1	–1
	4	–1	–1	–1	5	–1	–1	–1	–1	–1	5	–1	–1
	5	–1	–1	–1	–1	5	–1	–1	–1	–1	–1	5	–1
	6	–1	–1	–1	–1	–1	5	–1	–1	–1	–1	–1	5
l^{AB}	11	5	–1	–1	–1	–1	–1	–5	1	1	1	1	1
	12	–1	5	–1	–1	–1	–1	1	–5	1	1	1	1
	13	–1	–1	5	–1	–1	–1	1	1	–5	1	1	1
	14	–1	–1	–1	5	–1	–1	1	1	1	–5	1	1
	15	–1	–1	–1	–1	5	–1	1	1	1	1	–5	1
	16	–1	–1	–1	–1	–1	5	1	1	1	1	1	–5
	21	–5	1	1	1	1	1	5	–1	–1	–1	–1	–1
	22	1	–5	1	1	1	1	–1	5	–1	–1	–1	–1
	23	1	1	–5	1	1	1	–1	–1	5	–1	–1	–1
	24	1	1	1	–5	1	1	–1	–1	–1	5	–1	–1
	25	1	1	1	1	–5	1	–1	–1	–1	–1	5	–1
	26	1	1	1	1	1	–5	–1	–1	–1	–1	–1	5

Table B.13 a **values for** $(I \times J)$ **log linear analysis** (continued)

(3 × 5) table

		a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{31}	a_{32}	a_{33}	a_{34}	a_{35}
l^A	1	2	2	2	2	2	–1	–1	–1	–1	–1	–1	–1	–1	–1	–1
	2	–1	–1	–1	–1	–1	2	2	2	2	2	–1	–1	–1	–1	–1
	3	–1	–1	–1	–1	–1	–1	–1	–1	–1	–1	2	2	2	2	2
l^B	1	4	–1	–1	–1	–1	–4	1	1	1	1	4	–1	–1	–1	–1
	2	–1	4	–1	–1	–1	1	–4	1	1	1	–1	4	–1	–1	–1
	3	–1	–1	4	–1	–1	1	1	–4	1	1	–1	–1	4	–1	–1
	4	–1	–1	–1	4	–1	1	1	1	–4	1	–1	–1	–1	4	–1
	5	–1	–1	–1	–1	4	1	1	1	1	–4	–1	–1	–1	–1	4
l^{AB}	11	8	–2	–2	–2	–2	–4	1	1	1	1	–4	1	1	1	1
	12	–2	8	–2	–2	–2	1	–4	1	1	1	1	–4	1	1	1
	13	–2	–2	8	–2	–2	1	1	–4	1	1	1	1	–4	1	1
	14	–2	–2	–2	8	–2	1	1	1	–4	1	1	1	1	–4	1
	15	–2	–2	–2	–2	8	1	1	1	1	–4	1	1	1	1	–4
	21	–4	1	1	1	1	8	–2	–2	–2	–2	–4	1	1	1	1
	22	1	–4	1	1	1	–2	8	–2	–2	–2	1	–4	1	1	1
	23	1	1	–4	1	1	–2	–2	8	–2	–2	1	1	–4	1	1
	24	1	1	1	–4	1	–2	–2	–2	8	–2	1	1	1	–4	1
	25	1	1	1	1	–4	–2	–2	–2	–2	8	1	1	1	1	–4
	31	–4	1	1	1	1	–4	1	1	1	1	8	–2	–2	–2	–2
	32	1	–4	1	1	1	1	–4	1	1	1	–2	8	–2	–2	–2
	33	1	1	–4	1	1	1	1	–4	1	1	–2	–2	8	–2	–2
	34	1	1	1	–4	1	1	1	1	–4	1	–2	–2	–2	8	–2
	35	1	1	1	1	–4	1	1	1	1	–4	–2	–2	–2	–2	8

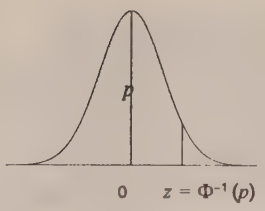
Table B.13 a **values for** $(I \times J)$ **log linear analysis** (continued)

(4 × 4) table

		a_{11}	a_{12}	a_{13}	a_{14}	a_{21}	a_{22}	a_{23}	a_{24}	a_{31}	a_{32}	a_{33}	a_{34}	a_{41}	a_{42}	a_{43}	a_{44}
l^A	1	3	3	3	3	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
	2	−1	−1	−1	−1	3	3	3	3	−1	−1	−1	−1	−1	−1	−1	−1
	3	−1	−1	−1	−1	−1	−1	−1	−1	3	3	3	3	−1	−1	−1	−1
	4	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	3	3	3	3
l^B	1	3	−1	−1	−1	3	−1	−1	−1	3	−1	−1	−1	3	−1	−1	−1
	2	−1	3	−1	−1	−1	3	−1	−1	−1	3.	−1	−1	−1	3	−1	−1
	3	−1	−1	3	−1	−1	−1	3	−1	−1	−1	3	−1	−1	−1	3	−1
	4	−1	−1	−1	3	−1	−1	−1	3	−1	−1	−1	3	−1	−1	−1	3
l^{AB}	11	9	−3	−3	−3	−3	1	1	1	−3	1	1	1	−3	1	1	1
	12	−3	9	−3	−3	1	−3	1	1	1	−3	1	1	1	−3	1	1
	13	−3	−3	9	−3	1	1	−3	1	1	1	−3	1	1	1	−3	1
	14	−3	−3	−3	9	1	1	1	−3	1	1	1	−3	1	1	1	−3
	21	−3	1	1	1	9	−3	−3	−3	−3	1	1	1	−3	1	1	1
	22	1	−3	1	1	−3	9	−3	−3	1	−3	1	1	1	−3	1	1
	23	1	1	−3	1	−3	−3	9	−3	1	1	−3	1	1	1	−3	1
	24	1	1	1	−3	−3	−3	−3	9	1	1	1	−3	1	1	1	−3
	31	−3	1	1	1	−3	1	1	1	9	−3	−3	−3	−3	1	1	1
	32	1	−3	1	1	1	−3	1	1	−3	9	−3	−3	1	−3	1	1
	33	1	1	−3	1	1	1	−3	1	−3	−3	9	−3	1	1	−3	1
	34	1	1	1	−3	1	1	1	−3	−3	−3	−3	9	1	1	1	−3
	41	−3	1	1	1	−3	1	1	1	−3	1	1	1	9	−3	−3	−3.
	42	1	−3	1	1	1	−3	1	1	1	−3	1	1	−3	9	−3	−3
	43	1	1	−3	1	1	1	−3	1	1	1	−3	1	−3	−3	9	−3
	44	1	1	1	−3	1	1	1	−3	1	1	1	−3	−3	−3	−3	9

Table B.14 Values for the inverse normal distribution

The tables contains the values of $z = \Phi^{-1}(p)$ for $p \geq 0.50$, with Φ^{-1} the inverse standard normal distribution function. For $p < 0.50$ the relationship used is $\Phi^{-1}(p) = -\Phi^{-1}(1-p)$.



p	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.50	0.000	0.003	0.005	0.008	0.010	0.013	0.015	0.018	0.020	0.023
0.51	0.025	0.028	0.030	0.033	0.035	0.038	0.040	0.043	0.045	0.048
0.52	0.050	0.053	0.055	0.058	0.060	0.063	0.065	0.068	0.070	0.073
0.53	0.075	0.078	0.080	0.083	0.085	0.088	0.090	0.093	0.095	0.098
0.54	0.100	0.103	0.105	0.108	0.111	0.113	0.116	0.118	0.121	0.123
0.55	0.126	0.128	0.131	0.133	0.136	0.138	0.141	0.143	0.146	0.148
0.56	0.151	0.154	0.156	0.159	0.161	0.164	0.166	0.169	0.171	0.174
0.57	0.176	0.179	0.181	0.184	0.187	0.189	0.192	0.194	0.197	0.199
0.58	0.202	0.204	0.207	0.210	0.212	0.215	0.217	0.220	0.222	0.225
0.59	0.228	0.230	0.233	0.235	0.238	0.240	0.243	0.246	0.248	0.251
0.60	0.253	0.256	0.259	0.261	0.264	0.266	0.269	0.272	0.274	0.277
0.61	0.279	0.282	0.285	0.287	0.290	0.292	0.295	0.298	0.300	0.303
0.62	0.305	0.308	0.311	0.313	0.316	0.319	0.321	0.324	0.327	0.329
0.63	0.332	0.335	0.337	0.340	0.342	0.345	0.348	0.350	0.353	0.356
0.64	0.358	0.361	0.364	0.366	0.369	0.372	0.375	0.377	0.380	0.383
0.65	0.385	0.388	0.391	0.393	0.396	0.399	0.402	0.404	0.407	0.410
0.66	0.412	0.415	0.418	0.421	0.423	0.426	0.429	0.432	0.434	0.437
0.67	0.440	0.443	0.445	0.448	0.451	0.454	0.457	0.459	0.462	0.465
0.68	0.468	0.470	0.473	0.476	0.479	0.482	0.485	0.487	0.490	0.493
0.69	0.496	0.499	0.502	0.504	0.507	0.510	0.513	0.516	0.519	0.522
0.70	0.524	0.527	0.530	0.533	0.536	0.539	0.542	0.545	0.548	0.550
0.71	0.553	0.556	0.559	0.562	0.565	0.568	0.571	0.574	0.577	0.580
0.72	0.583	0.586	0.589	0.592	0.595	0.598	0.601	0.604	0.607	0.610
0.73	0.613	0.616	0.619	0.622	0.625	0.628	0.631	0.634	0.637	0.640
0.74	0.643	0.646	0.650	0.653	0.656	0.659	0.662	0.665	0.668	0.671
0.75	0.674	0.678	0.681	0.684	0.687	0.690	0.693	0.697	0.700	0.703
0.76	0.706	0.710	0.713	0.716	0.719	0.722	0.726	0.729	0.732	0.736
0.77	0.739	0.742	0.745	0.749	0.752	0.755	0.759	0.762	0.765	0.769
0.78	0.772	0.776	0.779	0.782	0.786	0.789	0.793	0.796	0.800	0.803
0.79	0.806	0.810	0.813	0.817	0.820	0.824	0.827	0.831	0.834	0.838
0.80	0.842	0.845	0.849	0.852	0.856	0.860	0.863	0.867	0.871	0.874
0.81	0.878	0.882	0.885	0.889	0.893	0.896	0.900	0.904	0.908	0.912
0.82	0.915	0.919	0.923	0.927	0.931	0.935	0.938	0.942	0.946	0.950
0.83	0.954	0.958	0.962	0.966	0.970	0.974	0.978	0.982	0.986	0.990
0.84	0.994	0.999	1.003	1.007	1.011	1.015	1.019	1.024	1.028	1.032
0.85	1.036	1.041	1.045	1.049	1.054	1.058	1.063	1.067	1.071	1.076
0.86	1.080	1.085	1.089	1.094	1.098	1.103	1.108	1.112	1.117	1.122
0.87	1.126	1.131	1.136	1.141	1.146	1.150	1.155	1.160	1.165	1.170
0.88	1.175	1.180	1.185	1.190	1.195	1.200	1.206	1.211	1.216	1.221
0.89	1.227	1.232	1.237	1.243	1.248	1.254	1.259	1.265	1.270	1.276
0.90	1.282	1.287	1.293	1.299	1.305	1.311	1.317	1.323	1.329	1.335
0.91	1.341	1.347	1.353	1.359	1.366	1.372	1.379	1.385	1.392	1.398
0.92	1.405	1.412	1.419	1.426	1.433	1.440	1.447	1.454	1.461	1.468
0.93	1.476	1.483	1.491	1.499	1.506	1.514	1.522	1.530	1.538	1.546
0.94	1.555	1.563	1.572	1.580	1.589	1.598	1.607	1.616	1.626	1.635
0.95	1.645	1.655	1.665	1.675	1.685	1.695	1.706	1.717	1.728	1.739
0.96	1.751	1.762	1.774	1.787	1.799	1.812	1.825	1.838	1.852	1.866
0.97	1.881	1.896	1.911	1.927	1.943	1.960	1.977	1.995	2.014	2.034
0.98	2.054	2.075	2.097	2.120	2.144	2.170	2.197	2.226	2.257	2.290
0.99	2.326	2.366	2.409	2.457	2.512	2.576	2.652	2.748	2.878	3.090

Appendix C *Weights of commodities*

The information contained in this appendix constitutes detailed weights for the commodities required in determining the South African Consumer Price Index. The weights were determined by means of a market research investigation conducted in the twelve metropolitan areas during 1990. The weight for each commodity or service is noted for the low, middle and high income groups, as well as for pensioners and totalled (total column). The weights in each column add up to 100.

The information has been reproduced with the consent of the Central Statistical Service.

CONSUMER PRICE INDEX

DETAILED WEIGHTS

Group/Item	Income group				Pensioners
	Total	Lower	Middle	Higher	
ALL ITEMS	100.00				100.00
Lower income group	19.42	100.00			
Middle income group	24.09		100.00		
Higher income group	56.50			100.00	
Commodities	57.92	64.20	60.58	54.62	51.78
Services	42.08	35.80	39.42	45.38	48.22
1. FOOD	18.64	25.16	20.85	15.45	19.87
1.1 GRAIN PRODUCTS	2.92	5.25	3.24	1.98	2.84
White bread	0.68	1.22	0.85	0.42	0.56
Brown and wholewheat bread	0.39	0.96	0.38	0.20	0.40
Other bread and bread rolls	0.10	0.06	0.10	0.11	0.12
Cake flour	0.18	0.30	0.22	0.11	0.17
Bread flour	0.07	0.24	0.05	0.02	0.09
Breakfast oats	0.07	0.08	0.09	0.05	0.10
Corn flakes/breakfast cereals	0.18	0.10	0.20	0.21	0.23
Mealie meal	0.37	1.17	0.33	0.12	0.25
Mealie rice/samp	0.05	0.14	0.05	0.02	0.05
Rice	0.26	0.52	0.34	0.14	0.24
Spaghetti, macaroni and other pasta	0.12	0.09	0.15	0.11	0.15
Biscuits	0.15	0.16	0.18	0.14	0.19
Rusks	0.05	0.03	0.05	0.06	0.07
Cake, tarts, pies and other baked products	0.16	0.07	0.15	0.19	0.19
Other grain products	0.08	0.11	0.09	0.06	0.04
1.2 MEAT	6.17	7.67	6.83	5.38	6.38
Fresh - beef and veal	1.85	2.37	1.87	1.66	1.73
- mutton and lamb	1.42	1.60	1.76	1.22	1.66
- pork	0.31	0.23	0.35	0.32	0.38
- poultry (including frozen)	0.81	1.13	0.96	0.64	0.84
- boerewors	0.63	1.04	0.68	0.46	0.60
- other sausage	0.26	0.64	0.27	0.12	0.19
Bacon	0.13	0.08	0.13	0.15	0.16
Ham	0.07	0.03	0.06	0.09	0.09
Other cold meat	0.40	0.34	0.47	0.39	0.43
Other meat and meat products	0.31	0.22	0.29	0.34	0.30
1.3 FISH AND OTHER SEAFOOD	0.70	0.82	0.81	0.62	0.78
Fresh or chilled	0.24	0.32	0.34	0.17	0.27
Frozen	0.19	0.11	0.19	0.22	0.25
Smoked, dried and salted	0.02	0.03	0.02	0.02	0.03
Tinned fish	0.15	0.28	0.14	0.10	0.14
Fish paste	0.03	0.04	0.04	0.03	0.04
Other fish	0.07	0.05	0.07	0.08	0.07

Group/Item		Income group				Pensioners
		Total	Lower	Middle	Higher	
1.4	MILK, CHEESE AND EGGS	2.28	2.72	2.60	1.99	2.49
	Fresh milk	0.96	1.07	1.10	0.86	1.07
	Cheese	0.36	0.24	0.40	0.39	0.44
	Condensed/evaporated/sterilized milk	0.15	0.25	0.20	0.10	0.17
	Milk powder	0.14	0.23	0.16	0.10	0.17
	Other milk products (including creamers, whiteners and mixtures)	0.31	0.33	0.30	0.30	0.27
	Eggs	0.36	0.61	0.44	0.24	0.37
1.5	FATS AND OILS	0.87	1.29	1.05	0.65	0.89
	Butter	0.13	0.15	0.15	0.11	0.14
	Fats	0.01	0.03	0.01	0.01	0.01
	Cooking and salad oil	0.25	0.47	0.31	0.14	0.22
	Margarine	0.41	0.52	0.49	0.34	0.45
	Peanut butter	0.08	0.12	0.09	0.06	0.07
1.6	FRUIT AND NUTS	1.18	1.14	1.22	1.18	1.38
	Deciduous fruit	0.30	0.34	0.31	0.28	0.35
	Subtropical fruit	0.27	0.31	0.28	0.25	0.32
	Citrus fruit	0.14	0.23	0.17	0.10	0.16
	Other fresh fruit	0.05	0.02	0.04	0.06	0.07
	Canned fruit	0.06	0.03	0.06	0.07	0.07
	Dried/crystallised fruit	0.06	0.03	0.06	0.08	0.08
	Fruit juices	0.24	0.14	0.25	0.27	0.29
	Nuts	0.04	0.03	0.04	0.05	0.05
	Other fruit and nut products	0.02	0.01	0.01	0.02	0.00
1.7	VEGETABLES	1.88	2.52	2.14	1.55	2.15
	Potatoes	0.35	0.56	0.42	0.25	0.43
	Onions	0.17	0.30	0.21	0.11	0.19
	Tomatoes	0.26	0.39	0.28	0.20	0.29
	Green beans	0.07	0.09	0.08	0.06	0.08
	Cabbage	0.09	0.22	0.10	0.04	0.09
	Carrots	0.09	0.12	0.11	0.08	0.11
	Pumpkin and marrows	0.09	0.12	0.09	0.07	0.09
	Squashes	0.05	0.04	0.05	0.05	0.06
	Beetroot	0.05	0.08	0.06	0.03	0.04
	Sweet potatoes	0.05	0.09	0.06	0.03	0.05
	Lettuce	0.07	0.04	0.07	0.08	0.08
	Other fresh vegetable	0.09	0.05	0.09	0.10	0.09
	Dried/dehydrated vegetables	0.01	0.01	0.01	0.00	0.01
	Canned vegetables	0.20	0.27	0.24	0.16	0.19
	Frozen vegetables	0.24	0.13	0.27	0.27	0.34
	Other vegetable products	0.02	0.01	0.01	0.02	0.01
1.8	SUGAR	0.48	0.97	0.55	0.28	0.46
	White sugar	0.44	0.94	0.52	0.24	0.41
	Other sugar	0.04	0.03	0.04	0.04	0.05
1.9	COFFEE, TEA AND COCOA	0.59	0.89	0.66	0.45	0.70
	Coffee	0.28	0.38	0.31	0.23	0.34
	Tea	0.26	0.47	0.29	0.17	0.30
	Cocoa, chocolate drinks and other hot drinks	0.05	0.04	0.05	0.05	0.06

Group/Item		Income group				Pensioners
		Total	Lower	Middle	Higher	
1.10	OTHER FOOD PRODUCTS	1.38	1.40	1.55	1.29	1.44
	Salt	0.06	0.13	0.07	0.03	0.06
	Spices and flavouring	0.23	0.29	0.27	0.19	0.22
	Vinegar	0.04	0.07	0.05	0.03	0.04
	Yeast, baking powder and bicarbonate of soda	0.04	0.05	0.05	0.03	0.04
	Custard powder and puddings	0.05	0.06	0.05	0.05	0.05
	Canned soup, soup powder and stock cubes	0.06	0.06	0.06	0.05	0.08
	Chocolates	0.20	0.13	0.23	0.21	0.22
	Other sweets	0.05	0.02	0.04	0.06	0.05
	Ice cream	0.15	0.10	0.16	0.16	0.17
	Jam, syrup and honey	0.14	0.11	0.15	0.14	0.18
	Jelly powder	0.04	0.04	0.05	0.03	0.04
	Other related sugar products	0.01	0.01	0.01	0.01	0.01
	All other food products	0.32	0.32	0.37	0.30	0.30
1.11	FREE FOOD RECEIVED	0.19	0.49	0.22	0.08	0.37
2.	NON-ALCOHOLIC BEVERAGES	0.69	1.06	0.76	0.53	0.53
	Aerated	0.45	0.68	0.49	0.36	0.32
	Other non-alcoholic beverages	0.24	0.38	0.28	0.17	0.20
3.	ALCOHOLIC BEVERAGES	0.96	1.60	0.97	0.74	0.72
	Spirits	0.29	0.24	0.33	0.28	0.35
	Wine	0.15	0.08	0.14	0.18	0.15
	Beer (incl. sorghum beer)	0.52	1.26	0.50	0.28	0.21
	Other alcoholic beverages	0.01	0.02	0.01	0.00	0.01
4.	CIGARETTES, CIGARS AND TOBACCO	1.21	2.43	1.47	0.65	0.69
	Cigarettes	1.16	2.35	1.43	0.63	0.65
	Other	0.06	0.18	0.04	0.02	0.04
5.	CLOTHING AND FOOTWEAR	7.02	11.28	7.35	5.41	4.08
5.1	CLOTHING	5.87	9.12	6.08	4.66	3.41
5.1.1	WOMEN'S	1.72	2.21	1.91	1.47	1.29
	Coats and blazers	0.04	0.04	0.03	0.04	0.02
	Dresses and slack suits	0.68	0.93	0.80	0.54	0.62
	Slacks and skirts	0.18	0.15	0.19	0.19	0.12
	Blouses	0.17	0.23	0.15	0.16	0.10
	Jerseys	0.13	0.23	0.17	0.08	0.08
	Other outer clothing	0.13	0.06	0.11	0.16	0.08
	Dressing gowns	0.03	0.05	0.03	0.02	0.02
	Nightwear	0.08	0.11	0.11	0.07	0.05
	Underwear	0.16	0.22	0.18	0.13	0.12
	Stockings	0.08	0.11	0.09	0.06	0.07
	Other women's clothing	0.04	0.06	0.06	0.03	0.02

Group/Item	Income group				Pensioners
	Total	Lower	Middle	Higher	
5.1.2 GIRLS'	0.80	1.19	0.94	0.61	0.39
Coats	0.01	0.01	0.01	0.01	0.00
Dresses - school	0.08	0.15	0.09	0.05	0.03
- other	0.24	0.40	0.31	0.17	0.13
Blouses	0.07	0.08	0.07	0.06	0.04
Jerseys	0.06	0.13	0.07	0.03	0.04
Blazers	0.01	0.01	0.02	0.02	0.01
Other outer clothing	0.19	0.23	0.23	0.15	0.08
Underwear	0.06	0.09	0.08	0.05	0.03
Stockings	0.02	0.03	0.02	0.02	0.01
Nightwear and dressing gowns	0.05	0.06	0.05	0.05	0.02
Other girls' clothing	0.01	0.01	0.01	0.01	0.00
5.1.3 MEN'S	2.02	3.83	1.81	1.49	0.98
Coats	0.02	0.03	0.02	0.02	0.01
Lounge and dress suits	0.30	0.31	0.26	0.32	0.15
Safari suits	0.04	0.07	0.04	0.02	0.01
Jackets	0.13	0.20	0.12	0.12	0.05
Trousers - long	0.43	1.04	0.36	0.25	0.18
- other	0.10	0.17	0.09	0.07	0.06
Shirts	0.26	0.48	0.25	0.20	0.12
Jerseys	0.13	0.34	0.11	0.07	0.06
Other outer clothing	0.15	0.26	0.14	0.11	0.06
Underwear	0.10	0.21	0.10	0.06	0.05
Nightwear	0.03	0.05	0.03	0.02	0.01
Socks and stockings	0.06	0.14	0.06	0.04	0.04
Other men's clothing	0.27	0.54	0.24	0.20	0.19
5.1.4 BOYS'	0.74	1.25	0.77	0.54	0.29
Suits	0.05	0.06	0.05	0.04	0.02
Coats and blazers	0.08	0.15	0.09	0.06	0.03
Trousers	0.25	0.50	0.28	0.15	0.10
Shirts	0.11	0.17	0.12	0.09	0.06
Jerseys	0.04	0.10	0.05	0.02	0.02
Underwear	0.04	0.07	0.05	0.03	0.02
Nightwear	0.01	0.01	0.01	0.01	0.01
Socks and stockings	0.03	0.03	0.03	0.02	0.01
Other boys' clothing	0.12	0.17	0.10	0.11	0.04
5.1.5 INFANTS'	0.21	0.35	0.23	0.15	0.10
5.1.6 MATERIAL, KNITTING WOOL, HIRE AND MAKING-UP OF CLOTHING	0.38	0.28	0.41	0.40	0.35
Material	0.24	0.19	0.27	0.24	0.20
Knitting wool and yarns	0.06	0.03	0.06	0.06	0.10
Patterns, lace and sewing cotton	0.03	0.02	0.04	0.04	0.03
Cost of making-up and repairs	0.03	0.04	0.02	0.04	0.02
Hire of clothing	0.01	0.00	0.01	0.02	0.01

Group/Item		Income group				Pensioners
		Total	Lower	Middle	Higher	
5.2	FOOTWEAR	1.15	2.16	1.27	0.76	0.67
	Women's	0.43	0.79	0.44	0.29	0.28
	Girls'	0.19	0.35	0.23	0.11	0.08
	Men's	0.34	0.63	0.35	0.24	0.23
	Boys'	0.16	0.29	0.20	0.10	0.06
	Infants'	0.02	0.05	0.02	0.01	0.01
	Repair of footwear	0.02	0.05	0.02	0.01	0.01
6.	HOUSING	20.54	16.20	19.46	22.49	30.18
6.1	RENT	3.46	4.16	5.39	2.39	4.35
	House rent	1.61	1.73	1.97	1.42	0.79
	Flat rent	1.84	2.43	3.42	0.97	3.56
6.2	HOME-OWNER'S COSTS	15.54	10.21	12.58	18.63	24.21
	Interest	11.51	7.66	9.17	13.83	2.93
	Assesment rates	0.89	0.64	0.89	0.97	1.73
	Sanitary service	0.15	0.09	0.11	0.19	0.26
	Refuse removal	0.13	0.12	0.11	0.14	0.20
	Insurance of buildings	0.11	0.04	0.06	0.15	0.20
	Repairs and maintenance	1.23	0.31	0.62	1.81	2.27
	Cost of ownership	1.52	1.35	1.62	1.54	16.62
6.3	OTHER	1.55	1.82	1.50	1.47	1.63
	Water	0.88	0.74	0.93	0.90	1.14
	Boarding - hostels	0.38	1.02	0.37	0.16	0.18
	- holiday	0.29	0.06	0.19	0.41	0.31
7.	FUEL AND POWER	3.26	3.90	3.45	2.96	3.89
	Electricity	2.42	2.08	2.84	2.37	3.40
	Gas	0.17	0.19	0.14	0.18	0.08
	Petroleum products	0.20	0.80	0.14	0.02	0.15
	Methylated spirits	0.03	0.06	0.02	0.03	0.03
	Wood and coal	0.25	0.69	0.22	0.11	0.16
	Other	0.18	0.07	0.10	0.25	0.08

Group/Item	Income group				Pensioners
	Total	Lower	Middle	Higher	
8. FURNITURE AND EQUIPMENT	5.50	5.58	6.17	5.18	4.24
8.1 FURNITURE	2.47	2.65	2.69	2.32	1.54
Bedroom suites	0.47	0.79	0.68	0.28	0.24
Diningroom suites	0.25	0.32	0.39	0.17	0.18
Lounge suites	0.51	0.56	0.42	0.52	0.27
Kitchen sets and units	0.18	0.29	0.27	0.10	0.06
Loose items of furniture	0.44	0.39	0.46	0.45	0.35
Carpets	0.17	0.13	0.17	0.18	0.17
Other floor coverings	0.24	0.06	0.12	0.35	0.09
Other furniture	0.23	0.11	0.20	0.28	0.20
8.2 APPLIANCES	1.57	1.38	1.76	1.55	1.59
Electrical - Refrigerators	0.11	0.17	0.18	0.06	0.06
- Freezers	0.07	0.05	0.06	0.09	0.06
- Deep freeze/ refrigerator combination	0.08	0.11	0.12	0.05	0.16
- Stoves and ovens	0.26	0.22	0.20	0.29	0.12
- Washing machines	0.15	0.13	0.22	0.13	0.10
- Vacuum cleaners and polishers	0.07	0.05	0.08	0.08	0.07
- Sewing machines	0.13	0.04	0.11	0.16	0.28
- Other electrical appliances	0.57	0.43	0.60	0.60	0.48
Non-electrical - Stoves and heaters	0.03	0.10	0.04	0.01	0.02
- Sewing and knitting machines	0.03	0.04	0.05	0.02	0.06
- Other non-electrical appliances	0.06	0.04	0.09	0.06	0.18
8.3 OTHER HOUSEHOLD EQUIPMENT AND TEXTILES	1.46	1.55	1.71	1.32	1.11
Glass and plasticware	0.09	0.08	0.11	0.09	0.08
Metalware	0.07	0.06	0.08	0.06	0.07
Curtains	0.30	0.18	0.21	0.37	0.23
Blankets	0.37	0.64	0.53	0.21	0.17
Sheets, pillowcases and bedspreads	0.16	0.22	0.24	0.11	0.12
Mattresses	0.11	0.11	0.18	0.08	0.14
Other textiles	0.15	0.14	0.17	0.15	0.13
Gardening equipment	0.06	0.01	0.03	0.08	0.04
Other equipment	0.02	0.01	0.03	0.02	0.03
Light bulbs and plugs	0.07	0.04	0.08	0.08	0.07
Other	0.05	0.07	0.05	0.05	0.05

Group/Item		Income group				Pensioners
		Total	Lower	Middle	Higher	
9.	HOUSEHOLD OPERATION	2.57	3.27	2.32	2.44	1.96
9.1	HOUSEHOLD CONSUMABLES	1.37	2.16	1.45	1.07	1.02
	Soap	0.11	0.27	0.13	0.04	0.06
	Washing powder	0.32	0.65	0.34	0.19	0.22
	Liquid detergents and bleaches	0.21	0.26	0.26	0.17	0.17
	Brushes and brooms	0.02	0.01	0.02	0.02	0.01
	Floor polish and cleaning materials	0.09	0.20	0.13	0.04	0.06
	Shoe and other polish and cleaning materials	0.13	0.25	0.16	0.08	0.08
	Scouring materials	0.04	0.06	0.05	0.03	0.03
	Fertilizer and lime	0.06	0.01	0.03	0.08	0.06
	Disinfectants and swimming pool chemicals	0.20	0.10	0.15	0.26	0.16
	Matches and candles	0.09	0.29	0.08	0.02	0.05
	Other	0.11	0.06	0.11	0.13	0.11
9.2	DOMESTIC WORKERS	0.83	0.20	0.55	1.17	0.75
9.3	OTHER HOUSEHOLD SERVICES	0.37	0.91	0.32	0.20	0.19
	Laundry and dry-cleaning services	0.30	0.90	0.25	0.11	0.13
	Other household services	0.07	0.02	0.06	0.09	0.06
10.	MEDICAL CARE AND HEALTH EXPENSES	5.22	4.51	5.43	5.37	5.70
	Doctors' fees, nurses' fees and fees for related services	2.36	2.18	2.44	2.39	2.39
	Hospital, nursing-home fees and fees for related services	0.85	0.82	0.89	0.85	0.81
	Medical and pharmaceutical products	1.48	1.16	1.54	1.57	1.99
	Therapeutic appliances and equipment	0.39	0.26	0.41	0.42	0.40
	Contributions to aid funds	0.14	0.11	0.15	0.15	0.11

Group/Item	Income group				Pensioners
	Total	Lower	Middle	Higher	
11. TRANSPORT	14.43	8.52	13.38	16.92	10.74
11.1 VEHICLES	5.51	1.18	4.25	7.53	4.41
11.2 RUNNING COSTS	5.76	2.59	6.03	6.74	4.99
Petrol and diesel	3.21	1.43	3.28	3.80	2.64
Oil and grease	0.05	0.02	0.07	0.05	0.04
Tyres and tubes	0.41	0.18	0.40	0.49	0.25
Batteries	0.08	0.05	0.11	0.08	0.10
Spare parts	0.32	0.27	0.38	0.31	0.21
Repairs	0.26	0.19	0.37	0.24	0.35
Servicing, retreading, washing and related services	0.27	0.07	0.22	0.36	0.26
Licence and registration	0.16	0.07	0.19	0.18	0.18
Insurance	0.28	0.06	0.22	0.38	0.36
Parking fees	0.08	0.03	0.08	0.10	0.04
Other running costs	0.64	0.21	0.72	0.75	0.58
11.3 PUBLIC AND HIRED TRANSPORT	3.16	4.75	3.10	2.65	1.34
Bus	0.58	1.39	0.75	0.22	0.45
Train	0.28	0.72	0.39	0.08	0.11
Aircraft	0.35	0.16	0.55	0.33	0.40
Taxi and hired transport	0.82	2.26	1.13	0.19	0.22
Free transport	1.14	0.21	0.28	1.83	0.17
12. COMMUNICATION	1.61	2.05	2.06	1.27	1.73
Telephone rental and installation	0.58	0.78	0.76	0.43	0.61
Telephone calls	0.94	1.13	1.21	0.76	1.02
Postage	0.08	0.12	0.07	0.06	0.08
Other post office expenses	0.02	0.02	0.02	0.02	0.02
13. RECREATION AND ENTERTAINMENT	3.71	2.08	3.21	4.49	3.05
Radios, compact disk players, record players, and tape recorders	0.27	0.21	0.19	0.33	0.13
Television sets	0.30	0.33	0.27	0.30	0.15
Gramophone records and magnetic tapes	0.05	0.01	0.03	0.07	0.03
Television licenses	0.41	0.64	0.56	0.26	0.38
Musical instruments	0.05	0.01	0.02	0.07	0.02
Photography - equipment	0.11	0.01	0.05	0.17	0.08
- films and developing	0.11	0.03	0.09	0.15	0.09

Group/Item		Income group				Pensioners
		Total	Lower	Middle	Higher	
	Aircrafts, boats and camping equipment	0.09	0.01	0.02	0.15	0.08
	Sport equipment	0.08	0.01	0.05	0.12	0.10
	Hobbies, toys and games	0.09	0.05	0.08	0.11	0.04
	Admission fees to bioscopes and zoos	0.16	0.08	0.16	0.19	0.16
	Membership fees - scientific and other associations, clubs and libraries	0.10	0.03	0.08	0.14	0.11
	Pets - food/feeds	0.25	0.13	0.25	0.29	0.37
	- other expenses	0.20	0.05	0.18	0.27	0.20
	Plants, seeds and flowers	0.19	0.05	0.11	0.28	0.19
	Other	1.23	0.43	1.06	1.59	0.93
14.	READING MATTER	0.75	0.78	0.86	0.70	0.65
	Books	0.24	0.15	0.23	0.27	0.14
	Newspapers	0.33	0.50	0.42	0.23	0.30
	Magazines	0.19	0.12	0.21	0.20	0.21
15.	EDUCATION	1.76	1.38	1.68	1.92	0.92
	Tuition and attendance fees	1.54	1.26	1.37	1.70	0.80
	Other	0.22	0.12	0.31	0.22	0.12
16.	PERSONAL CARE	2.64	3.06	2.95	2.36	2.04
	Hairdressing and beauty care services	0.51	0.41	0.52	0.53	0.48
	Talcum, face and baby powder	0.05	0.05	0.06	0.05	0.05
	Lipstick	0.05	0.04	0.05	0.05	0.04
	Cologne and perfume	0.24	0.27	0.28	0.22	0.13
	Deodorant	0.16	0.26	0.17	0.12	0.13
	Toothpaste and toothbrushes	0.18	0.32	0.20	0.11	0.13
	Hair preparations	0.27	0.28	0.34	0.23	0.20
	Razors, blades, and shaving cream	0.13	0.15	0.14	0.12	0.10
	Toilet soap	0.17	0.30	0.20	0.11	0.12
	Skin creams and preparations	0.33	0.21	0.29	0.38	0.24
	Toilet paper	0.26	0.39	0.33	0.18	0.22
	Other	0.31	0.39	0.35	0.26	0.20

Group/Item	Income group				Pensioners
	Total	Lower	Middle	Higher	
17. OTHER GOODS AND SERVICES	9.49	7.04	7.63	11.12	9.03
Watches and personal jewellery	0.32	0.44	0.26	0.31	0.20
Umbrellas and sunglasses	0.08	0.04	0.04	0.11	0.04
Handbags, travelling bags and school bags	0.10	0.16	0.07	0.10	0.06
Smokers' requisites	0.02	0.03	0.02	0.01	0.01
Perambulators and push-carts	0.04	0.03	0.04	0.04	0.05
Stationery	0.09	0.06	0.08	0.11	0.07
Drinks and prepared food bought away from home	2.34	2.42	1.89	2.51	2.27
Package tours and hotel accommodation	0.96	0.30	0.54	1.37	1.06
Interest on loans and bank charges	0.77	0.56	0.39	1.01	0.35
Life insurance, endowment policies, annuities and mortgage debt insurance	1.87	1.07	1.76	2.19	0.96
Insurance of contents of dwelling	0.37	0.07	0.25	0.53	0.41
Membership fees (trade unions, staff and professional associations)	0.10	0.09	0.08	0.11	0.08
Legal and other fees for professional services	0.24	0.04	0.27	0.30	0.13
Funeral and all other expenses	2.18	1.74	1.95	2.43	3.35

AREAS

For the calculation of indices in respect of the twelve urban areas, separate weights are used for each item, group and area. The weights are based on the total expenditure of the households in each of the relevant urban areas. The total weight for each area is as follows:

Total	100.0
Cape Peninsula	16.5
Port Elizabeth 1/	5.3
East London	1.9
Kimberley	0.9
Pietermaritzburg	1.7
Durban	9.3
Pretoria 2/	13.4
Witwatersrand	39.9
Klerksdorp 3/	1.7
Vaal Triangle 4/	4.3
O.F.S. Goldfields 5/	2.3
Bloemfontein	2.8

1/ Port Elizabeth, Uitenhage, Despatch
2/ Pretoria, Verwoerdburg and Akasia
3/ Klerksdorp, Orkney, Stilfontein
4/ Vereeniging, Vanderbijlpark, Sasolburg
5/ Welkom, Virginia, Odendaalsrus

Appendix D *Time series data*

The first fifteen data sets in this appendix contain time series from official publications issued by the Central Statistical Service (Pretoria) and are published with authorization by this department. Time sereis TR16 to TR22 were obtained from Ivor Jones Roy & Co. Inc. The following are short descriptions of each of the time series.

TR 1: Total number of children born alive annually in a certain population group ; 1915 to 1990

TR 2.1: Total month-end stocks of grain sorghum (1 000 bags of 90 kg) ; January 1976 to December 1984.

TR 2.2: Total month-end stocks of grain sorghum (1 000 ton) ; January 1984 to February 1993.

TR 3: Annual rainfall (mm) - Cape Town ; 1873 to 1991.

TR 4: Annual gross value of wheat production (R million) ; 1960 to 1991.

TR 5: Annual production of dairy products (tons) and eggs (million dozen) ; 1960 to 1991.

TR 6: Monthly export value of gold (R million) ; January 1975 to September 1992.

TR 7: Electricity supplied for domestic use annually (million kWh) ; 1920 to 1989.

TR 8: Bank debits - monthly average (R million) ; January 1976 to February 1993.

TR 9: Total monthly value of wool exports (R million) ; January 1975 to January 1993.

TR10: Total annual number of road accidents ; 1935 to 1992.

TR11: Amount of maize sold monthly by producers (1 000 tons) ; January 1979 to July 1992.

TR12: Value of gold sold annually (R million) ; 1910 to 1991.

TR13: Quarterly number of employment opportunities in the mining industry; first quarter 1977 to third quarter 1992.

TR14: Total number of motor cycles and scooters sold monthly; January 1987 to July 1992.

TR15: Amount of airmail (thousand kg) transported annually by the domestic services of the South African Airways ; 1965 to 1985

TR16: Monthly Johannesburg Stock Exchange Index for manganese shares; January 1970 to May 1993.

TR17: Monthly Johannesburg Stock Exchange Index for coal shares; January 1970 to May 1993.

TR18: Monthly Johannesburg Stock Exchange Index for gold shares; January 1970 to May 1993.

TR19: Monthly Johannesburg Stock Exchange Index for financial shares; January 1970 to May 1993.

TR20: Monthly Johannesburg Stock Exchange Index for industrial shares; January 1970 to May 1993.

TR21: Monthly Johannesburg Stock Exchange Index for all shares; January 1970 to May 1993.

TR22: Monthly rainfall in the Katse dam catchment area (Lesotho Highlands Water Scheme) October 1930 to September 1986.

***TR1**: Total number of children born alive in a certain population group - 1915 to 1990*

Year:	1915	1916	1917	1918	1919	1920	1921	1922	1923	1924
Total	40 471	41 196	40 722	41 582	39 724	43 445	43 302	42 832	42 181	42 346
Year:	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934
Total:	43 411	43 876	44 347	44 813	46 219	47 534	46 423	44 944	44 519	44 878
Year:	1935	1936	1937	1938	1939	1940	1941	1942	1943	1944
Total	47 717	48 630	50 878	52 065	53 517	54 439	54 469	56 143	58 765	61 253
Year:	1945	1946	1947	1948	1949	1950	1951	1952	1953	1954
Total:	59 489	64 060	66 301	66 477	66 523	65 492	66 258	68 031	69 049	68 924
Year:	1955	1956	1957	1958	1959	1960	1961	1962	1963	1964
Total	70 322	70 469	73 301	75 270	76 016	76 300	75 725	76 581	75 569	76 781
Year:	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
Total	77 342	79 426	81 635	84 100	85 758	88 368	89 115	89 890	89 965	83 224
Year:	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
Total:	79 811	78 513	73 998	73 183	73 079	74 760	79 058	77 686	81 139	81 534
Year:	1985	1986	1987	1988	1989	1990				
Total:	79 863	72 955	70 341	69 189	70 964	69 649				

***TR2.1**: Total month-end stocks of grain sorghum (1 000 bags of 90 kg) January 1976 to December 1984*

Year	Jan	Feb	March	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
1976	1 681	1 295	1 011	833	818	1 721	2 348	2 290	2 163	1 931	1 664	1 577
1977	1 398	1 257	1 111	1 321	2 809	3 786	3 860	3 741	3 369	3 042	2 728	2 360
1978	2 076	1 860	1 407	1 075	3 576	5 360	5 557	5 610	5 238	4 848	4 321	3 815
1979	3 260	2 672	1 971	1 912	3 101	3 460	3 446	3 253	3 005	2 628	2 413	2 251
1980	1 892	1 536	1 447	2 471	7 188	6 812	7 824	6 411	5 765	5 064	4 652	4 118
1981	3 428	2 927	2 204	1 870	2 271	4 029	5 031	5 142	4 875	4 552	4 174	3 840
1982	3 506	2 938	2 360	2 092	3 484	3 784	3 595	3 350	3 116	2 883	2 471	2 226
1983	1 914	1 692	3 450	1 781	2 226	2 294	2 304	2 181	2 048	1 847	1 625	1 469
1984	1 302	979	1 191	1 848	3 128	3 339	3 395	3 328	2 938	2 760	2 560	2 426

***TR2.2**: Total month-end stocks of grain sorghum (1 000 tons); January 1984 to February 1993*

Year	Jan	Feb	March	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
1984	117	88	107	166	281	300	305	299	264	248	230	218
1985	201	172	131	250	570	613	611	580	532	464	366	299
1986	254	502	182	127	322	380	357	286	177	104	94	145
1987	89	78	61	127	337	372	333	292	254	204	165	137
1988	101	76	57	71	289	379	357	315	280	237	215	172
1989	169	142	116	92	154	260	354	362	338	325	291	238
1990	200	175	142	104	132	274	275	250	230	206	173	157
1991	136	116	97	71	130	163	185	202	182	163	144	129
1992	109	92	79	87	106	122	120	146	150	124	126	145
1993	116	96										

TR3: *Annual rainfall (mm) - Cape Town; 1873 to 1991*

Year:				1873	1874	1875	1876	1877	1878	1879
Rainfall (mm):				603.8	665.5	653.0	676.7	903.2	1 042.2	476.3
Year:	1880	1881	1882	1883	1884	1885	1886	1887	1888	1889
Rainfall (mm):	449.8	650.5	744.5	814.3	718.6	708.6	731.3	586.2	915.9	786.9
Year:	1890	1891	1892	1893	1894	1895	1896	1897	1898	1899
Rainfall (mm):	669.0	769.6	1 039.4	594.9	557.5	588.0	470.4	510.3	731.0	680.5
Year:	1900	1901	1902	1903	1904	1905	1906	1907	1908	1909
Rainfall (mm):	539.5	675.6	857.0	760.0	808.2	765.8	514.6	505.2	627.9	609.9
Year:	1910	1911	1912	1913	1914	1915	1916	1917	1918	1919
Rainfall (mm):	547.6	696.5	559.1	611.9	635.5	660.4	538.7	642.9	568.5	497.6
Year:	1920	1921	1922	1923	1924	1925	1926	1927	1928	1929
Rainfall (mm):	658.0	656.1	485.4	708.4	478.0	620.3	504.0	481.6	423.2	496.1
Year:	1930	1931	1932	1933	1934	1935	1936	1937	1938	1939
Rainfall (mm):	418.6	484.9	603.0	458.7	485.9	556.8	538.0	712.2	638.8	491.5
Year:	1940	1941	1942	1943	1944	1945	1946	1947	1948	1949
Rainfall (mm):	614.2	833.6	692.7	547.4	849.6	636.5	572.8	573.8	585.7	521.2
Year:	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959
Rainfall (mm):	675.6	712.3	627.4	618.4	836.3	608.5	735.4	856.7	489.1	626.8
Year:	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
Rainfall (mm):	427.4	509.3	734.1	470.1	479.3	523.5	493.3	500.0	649.8	464.6
Year:	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
Rainfall (mm):	670.4	437.5	463.2	347.8	682.6	559.2	587.4	751.2	402.2	407.3
Year:	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Rainfall (mm):	476.6	584.7	439.0	526.8	567.2	569.1	584.4	651.6	507.1	680.9
Year:	1990	1991								
Rainfall (mm):	603.4	632.2								

TR4: *Gross value of wheat production (R million); 1960 to 1991*

Year:	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
Value:	43.5	45.9	52.3	40.1	51.8	65.3	40.7	35.5	74.9	84.0
Year:	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
Value:	88.2	94.1	119.4	121.5	146.1	156.6	188.9	272.9	228.4	230.1
Year:	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Value:	385.7	313.8	556.1	705.0	480.9	690.2	534.9	864.6	1 257.3	1 218.3
Year:	1990	1991								
Value:	930.7	860.0								

TR5: *Annual production of dairy products (tons) and eggs (million dozen); 1960 to 1991*

	Fresh milk	Industrial butter	Industrial cheese	Condensed milk	Milk powder	Skim milk powder	Eggs
	Million litres	Tons	Tons	Tons	Tons	Tons	Million dozen
Year	TR 5.1	TR 5.2	TR 5.3	TR 5.4	TR 5.5	TR 5.6	TR 5.7
1960	1 055	42 088	14 810	29 680	3 045	2 776	111
1961	1 051	48 184	17 155	31 233	3 348	2 882	114
1962	1 055	47 060	15 795	28 136	3 440	2 655	118
1963	1 069	43 500	14 768	31 315	4 159	3 259	123
1964	1 047	40 407	14 220	37 700	5 029	3 436	125
1965	1 047	38 241	14 165	34 591	4 884	3 924	127
1966	1 042	37 941	14 166	39 550	5 667	4 414	129
1967	1 051	44 423	18 532	42 797	7 674	7 477	142
1968	1 086	52 633	20 673	37 954	6 409	12 798	144
1969	1 077	55 434	22 092	42 302	7 220	11 957	154
1970	770	47 267	17 605	36 412	6 069	10 496	163
1971	970	43 964	17 680	42 211	7 423	11 670	157
1972	943	45 419	19 250	42 960	7 602	13 312	164
1973	993	38 370	20 888	45 293	9 349	13 364	177
1974	1 000	31 609	19 430	44 904	7 794	13 934	200
1975	1 003	26 572	22 003	50 194	6 909	15 745	229
1976	1 071	26 163	29 534	53 740	9 544	18 914	243
1977	1 071	23 869	28 729	42 769	8 353	20 657	246
1978	1 072	16 962	28 505	43 195	11 193	14 513	250
1979	1 110	15 750	27 118	37 607	11 689	14 943	268
1980	1 087	18 751	33 528	41 642	8 563	20 133	238
1981	1 154	16 586	32 309	41 367	12 873	18 069	238
1982	1 141	14 532	34 675	36 913	13 512	16 517	264
1983	1 153	20 598	35 815	32 114	10 041	26 676	272
1984	1 155	18 147	36 156	30 743	12 317	23 824	272
1985	1 196	16 300	36 098	24 280	11 786	21 881	272
1986	1 105	17 536	33 846	22 996	11 370	24 540	273
1987	1 089	13 171	35 313	24 708	10 312	14 159	281
1988	1 225	11 918	40 336	23 740	8 555	12 961	310
1989	1 217	15 522	39 542	20 162	8 742	19 736	341
1990	1 216	16 651	42 437	22 709	10 477	20 864	346
1991	1 234	21 415	42 252	18 707	12 392	26 154	355

TR6: Monthly export value of gold (R million); January 1975 to September 1992

Year	Jan	Feb	March	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
1975	222.1	218.0	197.5	219.7	211.1	184.3	234.4	210.7	199.2	196.6	205.5	241.2
1976	218.0	160.5	219.9	201.9	160.1	241.0	185.0	183.1	198.9	160.1	197.6	220.3
1977	183.6	204.1	221.5	240.8	206.0	253.9	224.8	221.0	222.6	276.5	280.8	259.6
1978	306.5	255.6	267.5	270.1	330.2	320.8	350.9	375.5	288.8	442.4	431.2	285.0
1979	400.2	389.7	448.1	420.9	376.1	447.4	571.1	447.5	485.4	581.0	686.5	748.5
1980	913.9	1 035.7	845.2	694.9	565.0	711.2	1 065.4	871.5	947.8	836.8	916.6	736.6
1981	768.2	673.2	682.5	603.9	634.4	664.5	693.6	690.1	711.7	791.3	735.4	689.8
1982	499.3	630.4	612.9	588.5	551.0	631.6	747.3	699.8	923.3	997.4	954.0	791.3
1983	918.3	869.6	912.5	849.1	852.5	768.3	826.1	874.7	804.7	707.5	797.3	748.9
1984	1 009.4	820.2	925.1	777.9	866.4	907.4	1 036.4	1 007.6	906.6	1327.1	1 307.3	791.5
1985	1 344.2	1 113.9	1 105.0	1 070.4	1 162.3	923.1	1 278.5	1 469.6	1 332.2	1 858.8	1 547.3	1 261.7
1986	1 608.4	1 261.2	843.7	1 329.6	1 279.4	1 332.9	1 543.4	1 488.5	1 560.5	1 896.5	1 481.5	1 093.0
1987	1 645.8	1 412.9	1 362.0	1 545.9	1 336.7	1 230.1	2 106.5	1 413.9	1 084.5	1 852.0	1 474.2	1 343.0
1988	1 836.1	1 274.4	1 605.9	1 429.5	1 483.9	1 542.3	1 758.8	1 913.5	1 674.7	1 265.0	1 694.3	1 889.8
1989	1 712.0	1 158.7	1 875.5	1 616.6	1 367.4	1 553.4	1 654.3	1 741.5	1 521.8	1 822.5	1 777.5	1 427.5
1990	1 749.4	1 338.1	1 360.9	1 683.7	1 443.7	1 496.7	1 249.2	1 749.9	1 503.0	1 690.3	1 633.0	1 172.0
1991	1 887.1	1 269.3	1 174.7	1 814.4	1 576.8	1 606.8	1 831.2	1 848.7	1 341.6	2 250.4	1 561.4	1 485.7
1992	1 682.4	1 268.9	1 586.4	1 635.1	1 169.3	1 187.5	1 598.9	1 752.0	1 757.5			

TR7: Electricity supplied for domestic use (million kWh); 1920 to 1989

Year:	1920	1925	1930	1935	1940	1945	1950	1955	1958	1959
Amount:	32	96	142	336	690	927	1 563	2 553	2 784	3 078
Year:	1960	1961	1962	1963	1964	1965	1967	1969	1971	1973
Amount:	3 361	3 538	3 469	3 724	4 221	4 960	5 682	6 714	8 088	9 265
Year:	1975	1976	1977	1978	1979	1980	1981	1982	1983	1986
Amount:	9 991	10 474	11 302	11 718	12 330	13 522	14 738	14 618	14 490	17 644
Year:	1989									
Amount:	20 780									

TR8: Bank debits - monthly average (R million); January 1976 to February 1993

Year	Jan	Feb	March	April	May	June	July	Aug	Sep	Oct	Nov	Dec
1976	26 207	26 996	30 178	28 647	26 974	30 290	31 156	30 664	29 203	28 729	31 594	30 840
1977	28 197	27 960	32 175	30 592	29 979	32 922	34 856	34 515	34 326	31 675	32 698	34 636
1978	33 170	31 925	37 101	35 348	37 847	40 474	39 496	41 246	39 114	40 543	46 761	40 474
1979	41 470	41 924	43 677	42 518	44 651	49 391	49 821	52 078	47 834	58 704	55 495	52 921
1980	56 808	61 065	64 438	62 912	63 079	69 088	71 206	71 025	73 931	78 029	70 087	76 341
1981	73 168	75 445	79 141	76 441	78 304	83 287	90 252	87 068	90 313	90 146	82 851	93 209
1982	88 663	94 708	13 887	90 991	87 948	99 679	103 712	107 622	140 464	113 959	115 612	114 950
1983	119 429	110 198	143 819	113 568	114 762	136 733	120 677	130 662	177 900	133 257	129 896	150 424
1984	154 619	141 042	179 075	132 942	153 878	167 450	184 054	159 373	175 971	175 479	172 623	155 234
1985	166 563	156 711	172 990	155 554	162 568	165 693	187 874	186 482	155 582	168 889	175 737	167 893
1986	160 894	177 730	171 429	184 869	181 823	188 302	196 463	202 606	217 037	208 205	209 332	251 371
1987	249 779	245 592	275 816	274 938	244 425	288 863	306 744	324 290	352 223	353 289	332 450	348 234
1988	331 860	362 387	305 948	234 184	248 307	259 250	253 597	261 777	266 345	256 206	269 782	271 061
1989	256 906	269 471	315 798	292 512	296 128	324 580	307 810	344 485	335 047	340 659	344 531	353 295
1990	354 357	353 067	411 917	372 192	362 121	384 340	392 072	426 971	438 495	507 594	457 734	417 148
1991	450 112	414 626	436 562	486 112	455 744	469 239	458 319	489 893	445 578	524 234	501 126	489 741
1992	538 901	503 605	534 933	545 928	489 469	622 841	747 764	634 356	674 163	779 211	644 278	677 460
1993	604 939	627 458										

***TR 9**: Total monthly value of wool exports (R million); January 1975 to January 1993*

Year	Jan	Feb	March	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
1975	7.1	9.2	10.4	14.1	12.0	12.1	7.2	4.0	7.3	13.3	20.0	20.6
1976	11.1	18.7	24.4	20.3	17.5	19.8	8.8	8.6	12.6	21.6	20.2	20.0
1977	14.6	21.0	29.7	16.5	21.7	21.1	7.7	7.6	13.6	24.0	21.7	25.8
1978	16.3	19.6	21.6	22.9	18.3	19.8	11.3	11.1	18.1	23.6	25.0	23.5
1979	15.0	24.0	26.4	22.6	23.0	19.0	13.3	11.6	14.4	24.7	22.2	23.3
1980	23.2	30.6	31.8	21.4	19.6	16.9	11.3	13.5	17.9	24.3	22.5	28.1
1981	21.6	28.3	30.5	25.2	28.3	22.1	19.4	15.6	20.6	27.5	34.7	34.1
1982	24.5	34.0	42.5	29.0	29.8	28.2	17.8	15.7	20.9	20.9	25.3	30.0
1983	23.0	38.7	42.8	32.3	34.3	32.7	23.4	21.1	32.2	39.8	48.2	42.4
1984	37.4	57.3	58.1	47.8	42.2	37.9	31.1	26.7	37.8	63.9	71.3	69.6
1985	47.5	48.2	42.8	24.3	39.8	28.4	16.4	18.4	39.3	42.9	45.2	51.8
1986	38.0	36.0	40.6	31.1	31.1	30.0	28.4	17.6	26.1	35.5	41.0	40.5
1987	24.4	48.0	51.5	33.3	33.7	37.9	23.2	36.0	25.9	42.1	48.5	56.0
1988	40.2	63.9	80.4	70.6	53.6	88.0	34.0	52.6	53.4	105.8	63.0	120.7
1989	52.6	124.2	104.0	83.6	66.3	80.7	26.9	28.2	53.7	75.8	72.0	69.4
1990	57.3	99.0	73.7	87.8	87.6	56.2	22.1	35.5	28.8	39.2	52.7	62.7
1991	78.4	76.1	69.0	65.6	78.7	88.3	41.0	41.3	40.7	61.5	81.3	51.1
1992	49.4	63.6	68.0	57.1	45.1	57.9	39.7	25.3	49.7	44.4	58.2	53.3
1993	36.1											

***TR10**: Total annual number of road accidents; 1935 to 1990*

Year:					1935	1936	1937	1938	1939	1940
Total:					31 759	39 276	45 947	43 868	43 205	36 655
Year:	1941	1942	1943	1944	1945	1946	1947	1948	1949	1950
Total:	34 423	27 170	22 404	20 810	23 522	39 405	47 421	54 369	56 398	51 966
Year:	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960
Total:	55 633	58 443	59 320	70 181	77 479	80 263	93 500	101 895	108 321	116 688
Year:	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970
Total:	114 789	110 482	121 296	139 401	144 497	152 683	163 875	168 716	192 515	205 267
Year:	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
Total:	218 926	218 436	221 528	213 103	260 762	269 565	263 365	288 638	265 099	319 507
Year:	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Total:	379 852	392 971	388 599	412 097	369 185	372 668	387 148	418 483	434 935	433 287
Year:	1991	1992								
Total:	444 541	429 485								

***TR11**: Amount of maize sold monthly by producers (1 000 t); January 1979 to February 1993*

Year	Jan	Feb	March	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
1979	1	1	1	2	2 900	2 763	1 440	441	88	22	13	4
1980	1	0	1	0	4 013	3 445	1 739	568	186	18	3	2
1981	2	1	1	1	1 351	4 488	5 051	1 980	600	97	15	4
1982	1	1	1	1	1 286	3 353	2 237	576	49	7	2	1
1983	0	1	0	0	1 654	1 048	529	134	18	3	2	1
1984	0	1	2	16	1 316	1 077	659	298	40	9	5	2
1985	2	1	2	10	1 932	2 883	1 607	281	35	7	2	1
1986	1	0	5	91	1 326	2 792	2 611	451	42	7	2	1
1987	0	0	0	0	2 346	2 746	1 210	194	38	7	2	1
1988	1	1	1	0	1 428	2 389	1 631	578	72	11	3	2
1989	1	1	1	1	423	1 473	4 368	3 235	487	36	5	2
1990	1	1	1	0	201	2 039	3 543	1 210	80	13	4	2
1991	2	0	0	0	446	1 248	3 633	1 399	93	11	3	2
1992	1	1	1	0	1 010	873	332	46	24	8	8	2
1993	4	1										

***TR12**: Value of gold sold (R million); 1910 to 1991 (1990=100)*

Year:	1910	1915	1920	1925	1930	1935	1940	1945	1950	1955
Value:	64.0	77.3	91.2	81.5	91.0	153.1	236.0	210.6	289.6	365.5
Year:	1960	1965	1970	1975	1980	1981	1982	1983	1984	1985
Value:	536.0	766.5	830.3	2 560.4	10 369.3	13 668.8	14 502.0	16 256.5	19 062.1	25 883.1
Year:	1986	1987	1988	1989	1990	1991				
Value:	29 475.3	28 700.3	33 712.4	37 551.8	38 046.7	39 376.5				

***TR13**: Quarterly number of employment opportunities in the mining industry; first quarter 1977 to third quarter 1992*

Quarter	1	2	3	4
1977	706 614	712 006	713 117	663 668
1978	658 059	664 571	670 864	662 418
1979	689 770	692 209	686 248	697 741
1980	708 456	714 379	710 889	705 087
1981	705 010	728 506	724 604	709 612
1982	725 700	707 552	704 344	683 982
1983	695 976	706 933	696 435	690 750
1984	709 575	711 511	721 470	701 198
1985	708 809	724 587	752 879	745 894
1986	759 897	756 637	765 772	752 264
1987	763 003	763 319	742 733	713 657
1988	675 831	670 671	728 970	711 232
1989	702 746	706 810	720 556	702 382
1990	704 660	692 900	689 317	660 151
1991	651 843	653 134	639 535	608 486
1992	623 143	607 950	595 676	

***TR14**: Total number of motor cycles and scooters sold monthly; January 1987 to July 1992*

Year	Jan	Feb	March	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
1987	369	507	310	327	330	268	239	235	260	298	224	272
1988	244	263	132	209	242	176	131	328	228	304	308	206
1989	217	296	253	245	260	295	242	206	242	205	248	201
1990	631	695	594	753	459	640	804	690	1 311	589	660	702
1991	1 294	1 300	732	1 191	1 135	805	638	693	698	611	356	529
1992	477	1 099	510	624	762	504	737					

***TR15**: Amount of airmail (thousand kg) transported annually by the domestic services of the South African Airways; 1965 to 1985*

Year:	1965	1966	1967	1968	1969	1970	1971	1972
Amount:	1 752	1 858	2 019	2 219	2 426	2 566	2 493	2 579
Year:	1973	1974	1975	1976	1977	1978	1979	1980
Amount:	2 633	2 947	3 008	4 079	4 374	4 559	4 900	5323
Year:	1981	1982	1983	1984	1985			
Amount:	5 531	5 840	5 998	6 086	5 271			

TR16: *Monthly Johannesburg Stock Exchange Index for manganese shares; January 1970 to May 1993*

Year	Jan	Feb	March	April	May	June	July	Aug	Sep	Oct	Nov	Dec
1970	61	68	73	71	56	52	57	55	53	55	48	47
1971	54	60	64	66	71	70	73	72	68	55	60	69
1972	73	65	68	66	70	75	80	86	87	90	96	100
1973	107	103	96	100	101	114	111	112	104	113	96	91
1974	97	95	121	121	105	97	103	106	100	102	122	117
1975	123	134	136	135	148	169	170	155	164	165	173	171
1976	181	174	183	199	193	198	217	207	221	211	201	211
1977	205	212	197	201	231	205	221	215	217	216	208	190
1978	194	174	164	173	176	178	176	197	195	208	194	205
1979	225	233	236	243	247	247	250	267	294	290	325	332
1980	382	412	421	383	413	408	439	400	418	456	498	475
1981	401	432	489	480	507	482	522	551	521	537	499	518
1982	499	488	469	511	401	402	476	535	546	546	556	595
1983	629	568	601	514	524	489	477	480	455	439	424	405
1984	485	461	487	509	526	545	535	596	584	598	640	639
1985	724	766	773	829	963	991	998	1 153	1 139	1 181	1 125	1 118
1986	1 068	1 054	1 181	1 153	1 265	1 195	1 089	1 139	984	942	935	984
1987	914	1 111	1 096	1 026	1 089	1 026	1 125	1 265	1 392	1 230	1 209	1 300
1988	984	914	1 406	1 251	1 511	1 968	1 954	1 947	2 004	2 320	2 179	2 461
1989	2 953	2 953	3 164	3 375	3 234	2 918	2 953	3 515	3 094	2 847	2 517	2 179
1990	2 531	2 707	3 375	3 270	3 797	4 079	4 219	3 235	2 778	2 918	2 883	3 333
1991	3 094	2 953	3 094	3 235	3 235	4 114	4 254	3 868	4 043	4 360	4 395	4 219
1992	4 360	3 797	4 184	4 043	4 571	5 204	4 782	3 868	3 839	3 410	3 516	3 305
1993	2 813	2 531	2 567	3 164	3 024							

TR17: *Monthly Johannesburg Stock Exchange Index for coal shares; January 1970 to May 1993*

Year	Jan	Feb	March	April	May	June	July	Aug	Sep	Oct	Nov	Dec
1970	139	139	138	146	139	136	150	157	157	151	144	139
1971	144	151	154	158	160	163	164	157	151	136	140	158
1972	149	154	149	145	148	143	143	148	152	152	154	150
1973	149	147	145	139	141	143	141	139	150	163	157	181
1974	177	197	226	209	197	168	167	169	154	138	131	137
1975	141	137	167	179	201	215	243	210	193	215	237	221
1976	225	209	216	256	272	264	332	291	304	284	293	326
1977	332	345	353	371	358	349	380	374	361	384	386	418
1978	398	349	360	394	429	458	488	532	558	573	554	625
1979	659	772	810	772	843	773	882	891	997	972	952	1 112
1980	1 052	1 096	1 128	1 060	1 174	1 244	1 423	1 586	1 505	1 659	1 636	1 634
1981	1 511	1 382	1 609	1 704	1 698	1 564	1 634	1 712	1 821	1 844	1 772	1 788
1982	1 636	1 363	1 309	1 461	1 294	1 155	1 341	1 443	1 464	1 212	1 267	1 281
1983	1 362	1 281	1 200	1 350	1 367	1 326	1 280	1 346	1 286	1 204	1 065	1 157
1984	1 147	1 239	1 302	1 347	1 378	1 413	1 352	1 485	1 440	1 482	1 647	1 614
1985	1 673	1 658	1 539	1 678	1 977	2 062	2 035	2 278	2 229	2 139	2 354	2 412
1986	2 077	1 840	1 777	1 597	1 741	1 550	1 382	1 351	1 335	1 174	1 316	1 244
1987	1 446	1 423	1 318	1 265	1 242	1 112	1 164	1 258	1 310	1 152	1 002	979
1988	1 068	1 060	1 033	1 024	975	1 115	1 369	1 329	1 315	1 450	1 424	1 442
1989	1 546	1 644	1 845	2 096	1 909	2 105	2 207	2 252	2 403	2 245	2 177	2 144
1990	2 235	2 282	2 439	2 317	2 743	2 748	2 699	2 885	2 752	2 902	2 642	2 673
1991	2 562	2 793	3 182	3 632	3 803	3 770	3 831	3 770	3 865	3 820	3 817	3 656
1992	3 851	3 773	3 820	3 738	3 761	3 329	2 855	2 606	2 481	2 313	2 106	2 094
1993	2 210	2 417	2 292	2 439	2 536							

***TR18**: Monthly Johannesburg Stock Exchange Index for gold shares; January 1970 to May 1993*

Year	Jan	Feb	March	April	May	June	July	Aug	Sep	Oct	Nov	Dec
1970	69	73	75	80	74	71	72	76	74	78	84	80
1971	80	85	88	84	83	82	82	75	73	65	77	74
1972	89	90	91	95	102	116	136	128	121	122	118	125
1973	125	134	151	165	180	215	189	185	193	190	175	251
1974	305	326	422	365	332	330	413	410	353	425	381	390
1975	322	337	327	287	307	325	306	231	229	233	215	219
1976	221	225	196	212	219	186	164	125	154	157	182	178
1977	159	192	172	161	157	155	168	171	192	195	200	214
1978	221	203	202	191	207	220	260	244	255	257	224	256
1979	271	280	265	257	292	295	289	330	396	362	431	512
1980	500	547	479	503	571	640	697	762	857	858	808	728
1981	571	578	590	583	583	455	524	603	639	606	565	541
1982	521	457	429	412	398	351	414	533	603	657	738	908
1983	984	711	713	806	807	822	848	881	759	643	790	834
1984	795	967	965	973	996	973	885	962	974	995	1 024	945
1985	911	900	1 051	1 066	1 032	995	909	1 037	1 087	1 122	1 185	1 164
1986	1 286	1 197	1 251	1 146	1 142	1 242	1 386	1 693	1 845	1 764	1 954	1 908
1987	2 032	1 897	1 992	2 209	2 113	2 066	2 410	2 388	2 277	1 828	1 904	1 774
1988	1 302	1 212	1 351	1 207	1 246	1 289	1 331	1 241	1 266	1 343	1 343	1 301
1989	1 327	1 423	1 586	1 501	1 373	1 527	1 592	1 571	1 664	1 766	2 029	2 049
1990	2 184	1 919	1 960	1 738	1 638	1 522	1 571	1 674	1 603	1 429	1 267	1 201
1991	1 088	1 067	972	1 044	1 236	1 374	1 327	1 077	1 133	1 175	1 233	1 131
1992	1 246	1 239	1 118	1 017	1 090	1 097	1 069	938	902	816	862	799
1993	904	972	1 205	1 461	1 662							

***TR19**: Monthly Johannesburg Stock Exchange Index for financial shares; January 1970 to May 1993.*

Year	Jan	Feb	March	April	May	June	July	Aug	Sep	Oct	Nov	Dec
1970	260	243	250	271	218	182	208	215	199	178	159	169
1971	174	173	169	169	202	206	209	189	175	161	168	196
1972	205	205	240	254	290	274	275	311	320	317	325	335
1973	321	318	323	297	292	293	280	271	243	257	224	205
1974	227	225	239	227	207	181	173	156	145	143	158	161
1975	170	179	173	175	183	208	204	182	160	167	181	178
1976	182	170	173	179	181	175	170	154	164	152	154	158
1977	156	156	152	151	151	149	160	172	175	179	184	193
1978	195	189	185	199	203	219	227	229	230	244	237	246
1979	275	276	292	287	294	283	292	317	335	366	365	373
1980	383	434	416	410	419	446	474	493	481	502	478	478
1981	432	451	482	495	486	445	458	476	493	496	500	505
1982	507	476	423	451	414	399	456	493	585	610	687	711
1983	786	807	780	850	888	883	889	931	932	865	912	1 014
1984	978	1 007	1 056	1 066	994	945	861	881	854	891	983	927
1985	832	815	852	962	1 092	1 173	1 110	1 124	1 123	1 049	1 119	1 140
1986	1 198	1 171	1 177	1 115	1 154	1 140	1 158	1 221	1 278	1 265	1 233	1 202
1987	1 225	1 257	1 377	1 480	1 517	1 459	1 468	1 457	1 539	1 266	1 093	1 134
1988	1 106	1 046	1 102	1 096	1 079	1 089	1 128	1 081	1 179	1 211	1 186	1 257
1989	1 362	1 428	1 440	1 465	1 383	1 440	1 480	1 646	1 631	1 513	1 585	1 747
1990	1 804	1 791	1 872	1 732	1 910	1 867	1 965	1 888	1 824	1 757	1 781	1 942
1991	1 881	1 958	2 156	2 295	2 323	2 379	2 519	2 535	2 482	2 629	2 622	2 605
1992	2 738	2 799	2 817	2 791	3 108	2 917	2 779	2 803	2 996	2 861	3 051	3 191
1993	3 366	3 446	3 504	3 457	3 536							

TR20: *Monthly Johannesburg Stock Exchange Index for industrial shares; January 1970 to May 1993*

Year	Jan	Feb	March	April	May	June	July	Aug	Sep	Oct	Nov	Dec
1970	253	238	229	244	197	174	192	203	198	183	160	163
1971	167	166	165	187	196	198	203	186	171	155	164	182
1972	187	188	209	216	225	216	215	242	246	250	256	270
1973	267	265	273	262	270	286	281	277	247	268	227	219
1974	246	251	265	261	234	209	190	174	167	172	185	187
1975	196	206	201	196	203	233	229	215	189	200	210	213
1976	216	202	202	206	214	204	200	182	196	180	180	193
1977	189	188	177	178	182	182	189	202	206	207	204	219
1978	217	203	202	221	226	240	253	262	264	282	268	275
1979	304	307	321	324	324	313	324	353	386	418	430	456
1980	473	523	511	499	543	564	602	632	616	671	622	610
1981	561	579	628	650	637	589	615	650	668	684	686	695
1982	689	647	548	573	507	481	536	574	650	640	694	705
1983	783	781	790	835	895	904	865	867	886	827	836	910
1984	891	925	974	985	965	940	819	833	790	812	880	846
1985	779	773	810	896	965	980	958	957	965	940	1 030	1 067
1986	1 089	1 156	1 166	1 089	1 157	1 174	1 232	1 300	1 379	1 390	1 389	1 423
1987	1 528	1 627	1 740	1 840	1 874	1 923	2 152	2 209	2 257	1 708	1 460	1 443
1988	1 457	1 429	1 503	1 485	1 574	1 634	1 685	1 621	1 737	1 846	1 856	1 954
1989	2 139	2 238	2 460	2 581	2 415	2 621	2 688	2 828	2 741	2 569	2 569	2 790
1990	2 923	3 010	2 976	2 804	2 946	2 963	3 037	2 901	2 660	2 710	2 792	3 018
1991	2 882	3 218	3 389	3 542	3 608	3 817	4 085	4 114	3 974	4 263	4 220	4 170
1992	4 402	4 418	4 485	4 313	4 665	4 520	4 243	4 060	4 211	4 013	4 192	4 363
1993	4 523	4 487	4 426	4 371	4 425							

TR21: *Monthly Johannesburg Stock Exchange Index for all shares; January 1970 to May 1993*

Year	Jan	Feb	March	April	May	June	July	Aug	Sep	Oct	Nov	Dec
1970	200	201	193	207	170	153	166	175	173	166	150	154
1971	157	158	160	167	170	171	173	157	146	129	143	160
1972	171	173	189	199	215	218	235	244	241	240	240	252
1973	249	255	262	258	272	296	277	270	251	260	216	254
1974	294	306	358	330	293	272	303	297	257	290	280	279
1975	260	270	261	246	259	286	275	231	220	226	224	226
1976	231	223	210	223	232	210	206	174	190	183	192	201
1977	192	208	196	193	190	190	204	211	221	221	225	243
1978	243	226	226	232	244	260	285	284	295	312	284	310
1979	337	352	345	341	366	354	360	398	449	465	495	565
1980	559	616	565	562	620	677	730	776	827	848	791	746
1981	630	637	679	688	687	602	644	702	730	726	700	692
1982	671	613	526	540	492	452	528	612	688	708	781	879
1983	963	824	839	918	950	955	956	981	916	813	901	950
1984	936	1 034	1 053	1 062	1 046	1 029	925	982	966	994	1 041	984
1985	949	939	1 034	1 088	1 133	1 128	1 072	1 162	1 189	1 218	1 303	1 322
1986	1 400	1 394	1 464	1 358	1 420	1 500	1 613	1 862	1 920	1 898	1 999	1 981
1987	2 130	2 096	2 163	2 350	2 332	2 331	2 631	2 704	2 677	2 041	1 926	1 819
1988	1 568	1 518	1 680	1 605	1 676	1 755	1 812	1 724	1 849	1 968	1 961	1 990
1989	2 158	2 285	2 533	2 594	2 398	2 628	2 675	2 784	2 766	2 696	2 859	2 976
1990	3 194	3 084	3 257	3 032	3 188	3 077	3 152	2 993	2 744	2 667	2 601	2 720
1991	2 556	2 803	2 877	3 033	3 115	3 306	3 491	3 349	3 297	3 526	3 542	3 440
1992	3 605	3 597	3 550	3 454	3 732	3 655	3 431	3 152	3 211	3 017	3 192	3 259
1993	3 433	3 418	3 560	3 733	3 904							

TR22: *Monthly rainfalll in the Katse dam catchment area (Lesotho Highlands Water Scheme)*

Oct	Nov	Dec	Jan	Feb	March	April	May	June	July	Aug	Sep	Year
19.8	16.1	40.9	144.2	96.1	66.7	85.5	33.6	9.1	14.2	9.3	4.6	1930/31
14.9	41.5	33.6	46.8	113.6	96.5	32.0	11.4	6.4	3.2	3.1	3.2	1931/32
3.7	59.0	54.9	18.8	55.7	68.6	27.5	7.6	4.8	3.7	2.7	2.5	1932/33
2.5	136.4	192.8	262.7	156.5	114.2	46.4	11.1	5.6	26.9	33.2	13.0	1933/34
54.5	186.3	202.7	103.2	49.0	87.1	68.2	27.3	9.4	5.6	3.6	3.8	1934/35
8.5	17.1	31.5	36.4	41.4	74.1	34.7	24.9	15.4	4.9	3.1	2.2	1935/36
56.6	220.7	151.0	208.1	225.8	128.9	40.0	10.1	4.6	3.3	2.3	2.2	1936/37
32.9	22.8	40.6	129.1	159.0	58.8	76.1	38.7	33.5	15.3	57.4	32.4	1937/38
66.3	40.0	110.5	175.3	190.0	85.2	39.5	38.1	15.4	9.1	17.8	15.7	1938/39
71.2	146.9	85.0	69.0	74.8	79.2	77.3	63.9	22.9	6.9	3.5	38.2	1939/40
21.8	74.9	124.1	125.4	115.5	47.1	60.0	27.8	7.1	4.5	3.4	15.7	1940/41
86.6	41.3	7.6	113.7	102.9	137.1	82.7	22.8	7.9	4.4	31.9	17.5	1941/42
52.9	101.9	152.9	160.3	63.5	81.6	129.8	115.4	39.7	41.7	31.7	14.2	1942/43
177.8	204.7	173.7	114.9	127.7	69.9	19.2	8.9	33.9	15.9	5.4	56.1	1943/44
66.0	66.0	24.6	11.5	73.3	163.5	73.4	15.2	6.9	4.4	2.2	2.2	1944/45
2.2	3.4	5.8	64.5	67.6	70.6	33.4	19.2	10.0	3.5	3.3	2.4	1945/46
123.0	113.1	38.0	31.0	71.4	89.9	57.6	17.1	5.5	4.4	3.4	21.1	1946/47
61.5	80.8	97.3	88.9	64.1	207.0	131.1	27.9	8.9	4.6	3.3	2.3	1947/48
25.7	14.2	6.0	91.7	52.0	38.8	21.2	7.6	3.6	3.4	3.3	3.6	1948/49
40.0	56.4	96.9	95.4	78.2	151.8	108.1	46.0	14.3	26.3	67.1	34.4	1949/50
24.2	15.4	53.6	78.7	73.1	35.6	33.9	20.2	8.5	4.3	4.7	5.0	1950/51
110.7	54.3	30.6	76.1	125.4	78.0	30.2	10.1	4.3	12.2	12.7	8.5	1951/52
12.6	54.8	72.7	33.6	130.2	84.8	100.3	44.5	9.0	4.6	3.5	3.6	1952/53
56.2	49.9	57.9	77.6	66.4	57.2	21.1	10.0	6.6	3.4	3.3	3.3	1953/54
4.6	35.1	36.6	160.8	202.8	94.3	34.3	16.5	9.0	4.6	3.4	3.3	1954/55
29.4	86.1	122.0	57.2	116.8	110.8	38.6	35.9	16.7	4.4	3.3	4.4	1955/56
77.5	135.4	264.8	220.2	103.8	76.5	34.3	9.5	7.7	9.9	27.2	196.0	1956/57
259.2	149.4	62.7	116.8	65.9	32.1	65.5	45.3	14.0	5.4	3.3	18.5	1957/58
17.9	70.7	91.6	53.6	38.6	21.6	70.7	141.5	58.1	33.3	17.9	5.8	1958/59
88.3	104.3	141.4	95.8	102.9	121.3	87.6	28.1	6.9	5.5	23.0	27.0	1959/60
63.6	98.1	121.5	154.4	73.6	68.2	96.4	81.2	31.8	10.2	6.7	8.9	1960/61
6.8	125.2	107.1	76.6	133.9	111.0	67.0	22.2	6.8	3.3	3.3	3.5	1961/62
4.5	51.8	32.2	178.8	188.9	93.3	82.2	27.3	9.1	9.9	6.9	4.5	1962/63
28.3	82.3	77.5	120.0	77.0	132.2	80.9	21.2	18.7	10.7	7.5	7.9	1963/64
152.0	84.6	46.4	85.1	36.7	11.3	51.5	23.2	9.1	9.3	7.6	7.7	1964/65
8.0	58.9	31.1	159.9	124.3	34.9	9.1	5.4	3.4	3.3	2.4	2.4	1965/66
6.8	29.4	71.0	254.3	245.6	128.0	74.6	39.2	14.3	5.6	4.4	4.3	1966/67
4.6	69.0	68.4	19.1	8.7	29.9	30.2	75.7	11.3	9.6	5.5	8.5	1967/68
13.6	14.3	58.5	18.3	17.3	34.8	60.0	32.1	27.0	6.8	4.5	4.6	1968/69
41.2	35.3	53.8	45.5	49.4	6.9	3.8	3.3	3.3	3.5	4.0	22.5	1969/70
86.5	45.5	80.8	109.1	63.5	44.9	71.1	22.0	5.7	5.4	3.4	2.4	1970/71
2.4	11.4	27.6	113.4	117.9	169.4	29.1	30.2	7.9	3.4	2.4	2.5	1971/72
16.2	50.8	5.7	1.1	94.0	33.4	33.8	6.3	1.5	0.3	44.8	37.2	1972/73
32.1	30.7	75.6	126.5	123.6	32.1	39.9	14.8	12.9	4.8	9.8	10.8	1973/74
2.8	244.4	101.2	108.2	150.6	129.7	21.6	11.9	2.6	4.3	2.4	49.6	1974/75
67.8	239.2	114.3	163.4	181.8	248.0	118.1	34.5	23.9	6.8	4.4	33.5	1975/76
218.8	196.9	14.5	64.6	117.2	193.4	41.9	9.8	1.8	0.4	0.2	9.2	1976/77
115.7	36.2	25.7	189.8	51.0	59.5	212.0	13.4	2.7	1.3	2.6	61.1	1977/78
61.3	25.3	134.9	26.7	26.5	26.9	2.2	7.8	8.8	29.5	145.6	120.9	1978/79
111.5	54.8	109.6	28.5	77.0	25.1	7.2	1.7	2.4	0.3	0.3	30.6	1979/80
20.2	74.8	46.9	246.2	82.3	106.1	76.5	22.9	34.0	3.8	32.5	29.7	1980/81
7.9	53.0	105.1	12.8	9.9	13.9	107.5	16.6	2.7	2.6	1.5	2.4	1981/82
53.1	145.9	15.5	4.8	16.4	26.7	18.9	25.6	7.4	4.8	6.0	5.6	1982/83
35.6	89.9	113.1	161.1	6.2	21.8	16.2	19.5	1.7	1.4	1.7	19.6	1983/84
14.3	51.9	21.2	15.0	12.4	77.4	13.8	2.9	0.3	0.1	0.1	0.1	1984/85
106.7	111.7	242.7	20.5	81.3	11.7	43.7	5.5	12.5	2.7	19.5	31.0	1985/86

Data obtained from MBL Group BLM00292

Index

Absolute values 129, 590
Absolute differences 590
Acceptance area 410
Acceptance number 41
Acceptance sampling 41, 306
Action line 125
Addition rule 288, 289, 297
Adjusted coefficient of determination 193–194
Alternative hypothesis 406, 407, 438, 439 (*see also* Composite hypothesis)
Analysis of variance 508
  for completely randomized block design 510
  for multiple comparison of means 519
  models for, 533
  notation for, 510
  sum of squares in, 511, 516
  summary of, 533
  tables for, 513, 523
  test procedures for, 512
Annual data
  model for, 222
Approximation of probability 308–309
Area determination 323, 326
Area sample 34, 38
Arithmetic mean 99, 110 (*see also under* Relative prices)
Array *see* Data array
Assumption for grouped data *see under* Grouped data
Asymptotic 345
Asymptote 196
Axiomatic definition of probability *see under* Probability

Back-to-back stem-and-leaf display 153
Bar chart 71 (*see also* Combined bar chart, Divided bar chart, Three-dimensional bar chart)
Bartlett's test 520
Base period 261–263 (*see also* Base year)
  shifting the, 261
  weights for the, 263
Base year 251
Basic counting rule 302–303
Bayes' rule 299
Bell-shaped distribution 267, 345
Bernoulli trials 332
Bias *see* Sample bias
Bimodal data 105
Bimodal distribution 105
Binomial distribution 332–339
  cumulative distribution function of the, 336
  mean of the, 333
  normal approximation of the, 358
  parameters of the, 333
  probability expression for the, 335
  standard deviation of the, 333
  symmetry of the, 335
  tables of the, 336
Binomial experiment 332
Bivariate data set 164
Bivariate linear regression theory 474
Block 520
Block sum of squares 521 (*see also* Mean block sum of squares)
Boundaries of probability *see* Probability boundaries
Bowley's coefficient of skewness 147
Box Jenkins model 239
Box-and-whisker plot 155
Business cycle 221

Categorical data 7
  frequency distributions of, 70
  graphic representations of, 70
Categorical frequency distribution 71
Categories 74, 548
Causal relations 182–183
Cell 39, 526
Census 16
  problems with, 17–19
Central limit theorem 353, 356–358
Central moment 143
Central tendency 96 (*see also* Location)
  measures of, 99–106 (*see also* Arithmetic mean, Geometric mean, Median, Mode, Quantiles, Weighted arithmetic mean)
Centered moving average 230, 231

Charlier's check 144
Chebyshev's theorem 138
Chi-square distribution 362–365
  mean of the, 362
  normal approximation of the, 365
  standard deviation of the, 362
  tables of the, 362
Chi-square goodness-of-fit test 549
Class boundary 59, 61, 113, 115
Class interval 58 (*see also* Modal interval, Open class interval)
Class mark 58, 109
Class width 58
Classic definition of probability *see under* Probability
Closed question 43
Cluster sampling 33
Co-factor 526
Cochran's rule 28
Cochran's test 558, 611
Coefficient ϕ 556
Coefficient of determination 174, 175, 498, 499
Coefficient of partial correlation *see* Partial correlation coefficient
Coefficient of quartile variation 141
Column percentage table *see* Two-way column percentage table
Combined bar chart 81
Combination notation 306
Combinations 303, 306
Combinatorial analysis 302
Complement rule of probability *see under* Probability
Completely randomized design 510–516
  calculation formulas for, 513
  notation for, 510
  test procedure for, 512
Completely randomized experiment 512
Composite hypothesis 407
Composite indices *see* Composite price indices
Composite price indices 252 (*see also* Fixed weight composite price indices, Unweighted composite price indices, Weighted composite price indices)
  selection of items in, 261
Computer in statistical analysis 9
Computer programs 500, 534
Computer-generated graphic representations 77
Concave shape of scatter diagram *see under* Scatter diagram
Conditional probability 291–292
Confidence band 486
Confidence coefficient 387, 392
Confidence interval 387–393 (*see also* Lower-tailed confidence interval, Non-central confidence interval, Scheffé intervals, Upper-tailed confidence interval)
  for an individual value in regression 489
  for the median 620–621
  for the population correlation coefficient 495
  for the population intercept α 479
  for the population mean 387
  for the population proportion 390
  for the population slope β 482
  for the population variance 389, 392
  for the two sample case 398
  length of a, 392
Consistency 385
Consumer price index of South Africa 262–266
Contingency coefficient 562
Contingency table 547
  fourfold 2×2, 553
  $I \times J$, 559, 561
  multi-way, 547
  one-way, 547
  three-way, 548
  two-way, 548
Continuity correction 360
Continuous data 6, 58, 62
Continuous distribution 323, 342
Continuous stochastic variable 318
Contrasts 519
Control characteristic 39
Control chart 125
Control question 47
Convex form of scatter diagram *see under* Scatter diagram
Correlation 174, 177, 491, 615
Correlation coefficient 174–183 (*see also* Multiple correlation coefficient, Partial correlation coefficient, Rank correlation coefficient)
  confidence interval for the, 495
  hypothesis testing for the, 493, 497
Counting techniques 302
Covariance
  analysis of, 526

CPI *see* Consumer price index of South Africa
Critical value 410, 586, 591, 595 (*see also* Left critical value, Right critical value)
Cross-section data 55
Cumulative binomial distribution *see* Binomial distribution
Cumulative distribution 324–328
Cumulative distribution function 324
Cumulative frequency distribution 65, 111
Cumulative frequency polygon 66
Cumulative Poisson distribution *see* Poisson distribution
Cumulative relative frequency distribution 69
Cumulative relative frequency polygon 69
Curve fitting 168–170 (*see also* Least squares curve)
Customary point estimator 382, 383
Cyclic movement 221, 236

Data (*see* Bimodal data, Categorical data, Continuous data, Cross-section data, Dichotomous data, Discrete data, Economic data, Missing data, Multimodal data, Nominal data, Ordinal data, Primary data, Qualitative data, Quantitative data, Secondary data, Time series data)
  kinds of, 6
Data analysis 5 (*see also* Exploratory data analysis, Time series analysis)
Data array 56
Data coding 3, 4, 47
Data collection 3
Data editing 3
Data encoding *see* Data coding
Data set 1 (*see also* Bivariate data set, Multivariate data set, Trivariate data set, Univariate data set)
Data sources 8
Data values 100
Deciles 107
  properties of, 109
Decision rule 414
Deflation 269
Degrees of freedom 512
Density function 320, 342, 345
Dependent variable 169, 472
Descriptive statistics 5, 280
Determination
  coefficient of, 175
Deviation *see* Explained deviation, Mean absolute deviation, Standard deviation, Total deviation, Unexplained deviation
Dichotomous data 611
Discrete data 7, 105
  frequency distributions of, 70
  graphic representations of, 70
Discrete stochastic variable 318
  cumulative distribution function of, 324
  mean of a, 329
  standard deviation of a, 329–330
Distribution *see* Bimodal distribution, Binomial distribution, Bell-shaped distribution, Chi-square distribution, Continuous distribution, Cumulative distribution, Empirical distribution, Exponential distribution, F distribution, Hypergeometric distribution, Normal distribution, $n(\mu, \sigma^2)$ distribution, Poisson distribution, Probability distribution, $r(n)$ distribution, Standard normal distribution, Stochastic variable, Student's t distribution, t distribution, Theoretical distribution, Unimodal distribution
Distribution function 324
Distribution-free techniques 583
Divided bar chart 75
Domain 318 (*see also under* Stochastic variable)
Double sampling plan 41
Double-log transformation 186 *see also* Power curve
Dow-Jones industrial shares index 267
Drobisch price index 256
Dual question 44

Economic data 107
Edgeworth price index 259
Effects 566, 571, 572 (*see also* Interaction effects, Main effects)
Efficiency 384
Empirical determination of mathematical laws 169
Empirical distribution 317, 319, 320
Error sum of squares 511, 521, 529 (*see also* Mean error sum of squares)
Errors *see* Observation error, Type I error, Type II error
Estimate 381
Estimation 381, 620 (*see also* Confidence intervals, Inference, Interval estimate, Point estimate)
  interval, 387

of parameters 568
point, 382
Estimator 382
bias of an, 385
consistent, 385
efficient, 384, 519
unbiased, 384
Event 281, 284, 287, 288
certain, 287
independent, 295
mutually exclusive, 289, 297
rare, 337
statistically independent, 295–297
Exceedance probability 420
Expected frequency 549
Experimental error 509
Experimental unit 509
Explained deviation 176
Explained variation 176
Explanatory variable 169 (*see also* Independent variable)
Exploratory data analysis 150 (*see also* Tukey)
Exponential curve 183–184, 223 (*see also* Modified exponential curve, Single-log transformation)
Exponential distribution 342
cumulative distribution function of the, 342
density function of the, 342
properties of the, 344
Extended stem-and-leaf display 155
Extrapolation 169–170

F distribution 367–370
tables of the, 368
F test 603
Factor 526, 527, 547, 553
level of a, 526
Factor reversal test 271, 272
Factorial *see n* factorial
Factorial design 526–535
notation for, 528
Factorial notation 11, 304
Failure 332, 340
Fieldwork 42
organization of, 46
Fieldworker 46
Finite correction factor 398
Fisher's exact test 556
Five-figure summary 154–155
Fixed effect model 533
Fixed weight composite price index 255
Fixed weights 268
Forecasting 219, 222, 280, 284
Fourfold 304
Fourfold table 553
Freehand sketching 166–172
Frequency distribution 56, 57, 58, 317 (*see also* Two-way frequency distribution)
Frequency polygon 62, 64
Frequency table 57
Friedman test 604, 607

Games of chance 282
Gauss distribution 344
General linear model 500
General price index *see* Fixed weight composite price index
Geometric mean 105
of relative prices 260
Gompertz curve 197–198, 201
Goodman 564
Goodness-of-fit test 351 (*see also* Chi-square goodness-of-fit test, Kolmogorov-Smirnov test)
Gosset, W.S. 365
Graphic representation 62, 70, 77, 180 (*see also* Computer-generated graphic representation)
Grids 44
Grouped data 99, 109
arithmetic mean for, 110
assumption for, 109
central moment for, 143
measures of location for, 109
measures of spread for, 134
mode of, 114
raw moment for, 143
standard deviation for, 134
Growth curve 183, 196 (*see also* Gompertz curve, Linear growth curve, Logistic curve, Modified exponential curve, Polynomial, Richards family of growth curves)
determining parameters for, 200
Growth curve parameter *see under* Growth curve
Growth factor 105

h rankings
Friedman test for, 610
problem of, 607, 610

Heterogeneous population 25, 39
Highly significant 413
Histogram 62, 63, 96
Homogeneity 562
Hyperbolic curve 188, 223 (*see also* Inverse transformation)
Hypergeometric distribution 339–340
  mean of the, 340
  standard deviation of the, 340
Hypergeometric experiment 340
  stochastic variable of, 339, 340
Hypothesis 406, 407 (*see also* Alternative hypothesis, Composite hypothesis, Null hypothesis, Pairwise hypotheses, Simple hypothesis, Statistical hypothesis)
  choice of, 417
Hypothesis testing 405 (*see also* Statistical hypothesis testing; *see also under* Coefficient of determination, Goodness-of-fit test, Homogeneity, Independence test, Linear regression, Log-linear analysis, Median, Proportion, Spread)
  differences between means for small samples, 430–432
  for bivariate samples 497
  for correlation coefficient, one sample, 493
  for correlation coefficients, two samples, 497
  for parameters for two populations 428
  for the population median, one sample 584
  for the population median, two samples 594
  guide for, 452
  large-sample, 421, 428–431
  left one-sided, 408
  left tailed, 408
  lower one-sided, 408
  lower-tailed, 408, 438
  mean, one-sample, *n* large, 421
  mean, one-sample, *n* small, 422
  means, paired observations, 433
  means, two dependent samples, 434
  means, two independent samples, *n* large, 428
  means, two independent samples, *n* small, 430
  one-sided, 425
  one-tailed, 425
  operating curve for null, 412
  p-values in, 419
  right one-sided, 408
  right tailed, 408
  small-sample, 422–424, 431–433
  steps in, 416
  two-sample, 419, 428
  two-sided, 408
  two-tailed, 408
  upper one-sided, 408
  upper-tailed, 408, 438
  variance, one sample, 426
  variance, two independent samples, 443
Ideal operating curve 412
Independence model 566–568
  estimating parameters for the, 568
  log linear representation of the, 567
Independence test 553–564
  $I \times J$, 559, 560
Independent samples 429, 431, 434, 601, 604
  difference in medians of, 594
  Kruskal-Wallis test for, 604
  rank sum test for, 594
Independent trials 332
Independent variable 169, 472 (*see also* Explanatory variable, Predictor)
Index chains 268–269
Indices 107, 250 (*see also* Composite indices, Consumer price index of South Africa, Dow-Jones industrial shares index, Johannesburg Stock Exchange indices, JSE overall index, Price indices, Quantity indices, Retail price indices, Seasonal indices, Simple composite price index chains, Simple indices, Value indices, Weighted composite price index chains)
  determination of, 260
  interpretation of, 260
  mathematical properties of, 271
Inference 360, 362, 364 (*see also* Estimation, Hypothesis testing, Statistical inference)
Inference assumptions
  for analysis of variance 512
  for linear correlation 491
  for linear regression 477
Inflation 250, 266
Influential observations 180, 182
Integration 331
Intelligence quotient 267
Interaction 450, 527
Interaction effect 566
Interaction sum of squares 529
Intercept 165

inferences regarding the, 477
least squares values of the, 170
Interpolation 169–170 (*see also* Linear interpolation)
Interquartile range 127
Interval estimate 286, 381–393, 479, 482, 489, 620
Interval estimation 387–393 (*see also* Confidence interval)
Interviewing 46
Inverse distribution function 349
Inverse transformation 188 (*see also* Hyperbolic curve)
Inverse normal distribution function 349
Inverse normal tables Table B.14
Inverse transformation 188
Inverse distribution function 349
Irregular variation 222
Irving Fisher price index 256, 272
*i*th percentile 111

Johannesburg Stock Exchange indices 267
JSE overall index 267

Kendall's coefficient of concordance 612
Kendall's coefficient of rank correlation 619
Kerrich, J.E. 282
Kolmogorov-Smirnov test 449, 553
Kruskal-Wallis test 603
Kurtosis 96, 98, 123, 147
coefficient of, 147, 148
measures of, 146

Laplace 344
Laspeyres price index 255, 257, 267
Laspeyres quantity index 259
Law of total probability 297
Least squares curve 168, 171, 173
Least squares method 167, 223–224
Least squares residuals 180
Left critical value 586, 591, 595
Left one-sided
confidence interval 392
test 408
Legendre, A. 168
Level of significance 411, 413
Linear correlation 174, 176, 177 (*see also* Inference assumption for linear correlation)
negative, 493, 494
Linear equation 170
Linear interpolation 350
Linear regression 472
bivariate, 474
confidence interval for, 480, 484
hypothesis testing for, 480, 484
Linear relation 170
strength of, 177
Location 96, 97, 99
measures of, 99–106, 584 (*see also* Arithmetic mean, Geometric mean, Measures of central tendency, Median, Mode, Quantiles, Weighted arithmetic mean)
Log linear analysis 564
Log linear model 565
effects of the, 566
parameters of the, 566
two-way, 566
Logistic curve 197–201
Long-term movement 220, 223–224
Long-term trend *see* Long-term movement
Lower one-sided test 408
Lower-tailed confidence interval 392
Lower-tailed testing 408
Lurking periods 32

m-folds 304
Main effect 566
Mathematical model 166
Mathematical notation *see* Notation
McNemar's test 553, 556–557
Mean (*see also* Moving average, Population mean; *see also under* Confidence intervals, Hypothesis testing)
arithmetic, 99
testing differences between, large samples 428
testing differences between, small samples 431
distribution of the, 365
geometric, 105, 106
hypothesis testing for, 421, 423, 429, 431, 433, 435, 590
population, 329, 420
weighted arithmetic, 192
Mean absolute deviation 129
Mean difference 398, 428, 431, 432, 433
Mean error sum of squares 512, 522, 525
Mean squared error 385
Mean treatment sum of squares 512, 522
Measure of dependence 615
Measurement interval 146, 337, 342

Median 103, 111, 584, 620
confidence interval for the, 620
hypothesis testing for the, 584, 589, 594, 597, 604
properties of the, 104
Midpoint 97
Missing data 534
Modal interval 114
Mode 104, 114
properties of the, 105
Model 256, 320, 402, 434, 463 (*see also* Stochastic effect model)
Modified exponential curve 196
Moments 143
Monotone relation 616, 619
Monthly data 222
Monthly indices 236
Movement component 219
Moving average 222, 227 (*see also* Centered moving average)
Multi-directional contingency table 547, 576
Multi-factor experiment 527
Multimodal 105
Multimodal distribution 105
Multiple-choice question 44
Multiple comparison technique 519
Multiple correlation coefficient 207
Multiple sampling plans 41
Multiplicative model 219, 222
Multi-sample problem 603
Multistage experiment 302–303
Multistage sampling 35
Multivariate data set 164
Multiway stratification 29

$n(\mu, \sigma^2)$ distribution 345
n factorial 11, 304
Natural logarithm 11–12
Nominal data 7, 105, 583
Non-central confidence interval 392
Non-probability procedures 38–40
Non-response 20
Non-sampling errors 20
Normal approximation of the binomial distribution 358
Normal approximation of the chi-square distribution 365
Normal approximation of the Poisson distribution 362
Normal approximation of the rank sum test 595
Normal approximation of the rank sign test 591
Normal approximation of the Siegel-Tukey test 601
Normal approximation of the t distribution 367
Normal approximation of the sign test 585
Normal curve 345 (*see also* Bell-shaped distribution)
Normal distribution 344–353
characteristics of the, 345
density function of the, 345
tables of the, 346–348
Normal equations 170
Normal probability distributions 360
Normal probability paper 553
Normal probability plot 351
Normality
assumption of, 451, 452
Notation 99 (*see also* Factorial notation, Combination notation, Set notation, Sigma notation, Summation notation; *see also under* Stochastic variable)
Null hypothesis 406, 407, 416

Observation errors 19 (*see also* Sampling observation errors)
Open class interval 58, 115
Open question 45
One-sample problem 370, 584
One-stage experiment 302
One-stage sampling problem 296, 346, 481
Open class interval 58
Open question 45
Operating curves 412 (*see also* Ideal operating curve)
Ordinal data 7, 583
Outcome 286
dichotomous, 332
Outlier 58, 155, 180, 182
Oversampling 39

p-value 420, 585, 591, 595, 601 (*see also* Exceedance probability)
Paasche price index 255
Paired equations 518
Paired observations 435, 598
Paired values 433, 434, 525
Pairwise comparison 517, 525, 532, 533 (*see*

also Scheffe's method of pairwise comparison, Tukey's method of pairwise comparison)
Pairwise comparison of means 517–519
Pairwise hypothesis 509
Parabola 192–194, 206, 223
Parameter 196, 566
Partial correlation 206
Partial correlation coefficient 206
Peakedness 96, 143
Pearson coefficient of skewness 147
Percentile 107, 111 (*see also* ith percentile)
Permutation 303–305, 597
Pie chart 71 (*see also* Three-dimensional pie chart)
Pilot study 39, 47
Pivotal quantity 358, 360, 364, 477, 482, 495
Point estimate 381
  for the population mean μ 382
  for the population proportion π 383
  for the population variance σ 383
  of the median 620
Point estimators 382
  characteristics of, 384
  customary, 382, 383
  table of, 399
Poisson distribution 336–339, 342
  normal approximation of the, 362
Poisson probability model 337
Polynomial of degree q 192
Population 16, 18, 100, 317 (*see also* Target population)
Population correlation coefficient 491, 615
Population elements 16
Population intercept 479
Population mean 100, 344
  confidence interval for the, 387
  testing for the difference between, dependent samples, 433
  testing for the difference between, large samples, 429
  testing for the difference between, two small samples, 430, 432
  testing for the, large sample case, 421
  testing for the, small sample case, 423
Population median 620
  testing for the, one sample 584
  testing for the, two samples 594
Population parameter 381, 406
Population proportion π 283
  confidence interval for, 390
  testing for, one sample, 424
  testing for, two samples, 440
Population regression line 474
Population regression plane 498
Population size 255
Population slope β 482
  inferences regarding, 482
Population standard deviation σ 330
  confidence interval for the, 389
  testing for the, one sample, 426
  testing for the, two samples, 442
Population variance σ^2 426
  confidence interval for the, 389
Positive skewness 350, 362
Positively skewed distribution 352
Positively skewed exponential distribution 357
Power curve 186 (*see also* Double-log transformation)
Power of a test 413, 443, 444, 445, 622
Predetermined value 406
Prediction 489
Prediction band 489
Predictor 169
Price indices 194 (*see also* Consumer price index, Drobisch price index, Edgeworth price index, Fixed weight composite price index, Laspeyres price index, Paasche price index, Simple composite price index, Simple composite value index, Unweighted composite price index, Weighted composite price index)
Prices *see* Relative prices
Primary data 8
Primary data sources 14
Probability 280–372 (*see also* Conditional probability, Total probability)
  axiomatic definition of, 287
  classical definition of, 284
  complement rule of, 289
  empirical method of determining, 284
  frequency limit definition of, 281–288, 330
  product rule of, 293
Probability boundaries 288
Probability distribution 320, 333, 334
Probability model 320
Probability sampling procedures 21 (*see also* Cluster sampling, Non-probability procedures, Stratified random sampling, Systematic sampling)

Probability rule 288–302 (*see also* Addition rule, Bayes' rule, Boundaries of probability, Complement rule, Conditional probability, Law of total probability, Product rule of probability, Statistically independent events)
Probability sample 21
Probability theory 5, 280–316
Process average 125
Product rule of probabilities 293–295
Properties of shape 96
Proportion
  confidence interval for the population, 390
  testing for the, one sample, 424
  testing for the, two samples, 440

Qualitative data 7, 72
Quality control 124
Quantiles 107, 111
  properties of the, 109
Quantitative data 7, 583
Quantity indices 252, 259 (*see also* Laspeyres quantity index, Value indices)
Quarterly data 222, 236
Quartile 107
Quartile deviation 127, 134
  properties of the, 128
Quartile variation 141
  coefficient of, 141
Questionnaire Appendix B, Data set 2
  appearance and layout of a, 45–46
  designing a, 42
Quota sampling 39–40 (*see also* Stratified random sampling)

Random block experiments 607
  Friedman test for, 607
Random digit 22
Random number 22
Random sample, 355 (*see also* Simple random sample)
  arithmetic mean of a, 357
Randomized block design 520–526
  calculation formulas for, 522
  notation for, 521
Randomness 612
Range 124, 134 (*see also* Interquartile range)
  properties of the, 125
Rank 589, 594, 600
Rank correlation 615
Rank correlation coefficient 615 (*see also* Spearman's rank correlation coefficient)
Rank sign test 589, 598 (*see also* Wilcoxon rank sign test)
  normal approximation of the, 591
  tables for the, 591
Rank sum test 594, 604 (*see also* Wilcoxon rank sum test)
  tables for the, 595
Ranking 590, 610
Rankits procedure 553
Raw data 56 (*see also* also Ungrouped data)
Raw moment 143
Real terms 269
Realization
  of an event 281, 284
Regression 472
Regression line 475 (*see also* Population regression line, Sample regression line)
  sample variance around the, 475–477
  variance around the, 379
Regression theory *see also* Linear regression theory
Rejection area 410
Rejection number 41
Relative efficiency 622
Relative frequency 66, 281, 284
Relative frequency distribution 66
Relative measures of spread 139, 142
Relative prices 254
  arithmetic mean of, 254
  geometrical mean of, 260
  weighted arithmetic mean of, 256, 264
Repetitions per cell 527
Residuals 180 (*see also* Least squares residuals)
  graphic representation of, 180–182
Richards family of growth curves 198
Right critical value 586, 591, 595
Right one-sided
  confidence interval, 392
  test, 408
Risk
  in hypothesis testing 411, 413
r(n) distribution 492
Row percentage table 74
Rules of probability *see* Probability rules
Runs test 612

Sample 16, 99 (*see also* Area sample, Cluster sample, Random sample, Simple random sample)
  acceptance, 41, 306

associated, 525
associated, Friedman test for, 607
cluster, 32–34
convenience, 39
dependent 434
dependent, difference in medians of, 597
judgmental, 39
multi-stage, 34
non-probability, 39
observed random, 355
one-stage, 35
probability, 351
proportionate, 25, 26
sequential, 42
stratified random, 25
systematic, 30–32
two-stage cluster, 38
Sample bias 20, 385
Sample moments 143 *see also* Central moments, Raw moments)
Sample ratio 30
Sample regression line 474
Sample size 19
determination of, 393
for estimating a population mean 394
Sample space 286, 287 (*see also* Certain event)
Samples *see* Dependent samples, Independent samples
Sampling 19 (*see also* Acceptance sampling, Double sampling, Multiple sampling, Oversampling, Quota sampling, Sampling with replacement)
arithmetic mean for, 355
standard deviation for, 355
Sampling distributions 355–372
one-sample problem 370
pivotal quantities for, 370, 371
Sampling error 20
Sampling methods 21 (*see also* Probability procedures, Non-probability procedures)
Sampling with replacement 333
Saturated model 566–569
log linear representation of the, 566
parameters for the, 568, 569
Saturation value 196
Scale of preference 45
Scatter diagram 99, 165, 166
concave form of a, 352
convex form of a, 352
Scatter plot *see* Scatter diagram
Scheffe's method of pairwise comparison 517, 532
Seasonal fluctuation 221, 230–236
Seasonal index 222, 230, 231 (*see also* Specific seasonal index, Typical seasonal index)
Seasonally adjusted time series 236
Secondary data 8
Semantic differential scale 45
Series *see* Time series, Weight series
Set notation 286
Set theory 286
Sheppard's correction 135
Siegel-Tukey test 600
normal approximation of the, 601
Sigma notation 9–11
Sign test 584–589
left critical value of the, 586
normal approximation of the, 585
right critical value of the, 586
tables of the, 585
Sign test statistic 585
Simple composite price index 253
Simple composite value index 259
Simple hypothesis 407
Simple indices 252
Simple random sample 22
Simple value index 252
Simulation 353
Single sampling plan 40
Single–log transformation 183
Skewness 98, 143, 359, 584 (*see also* Bowley's coefficient of skewness, Pearson coefficient of skewness)
coefficient of, 147
negative, 98
positive, 63, 98
Slope 165, 484
least squares values of the, 170
Sources of variation *see under* Variation
Spearman's rank correlation coefficient 615
Specific seasonal index 230
calculation of, 231
Spread 96, 97, 123, 139, 600 (*see also* Population spread, Quartile deviation, Range, Spread, Standard deviation)
difference in, 600
hypothesis testing for, 600
measures of, 123–124, 134–135
relative measures of, 139–141
Standard deviation 129–130, 134, 334, 337 (*see also* Population standard deviation)

confidence interval for the, 389
interpretation of the, 137
properties of, 133
testing for the, one sample, 426
testing for the, two samples, 442
Standard error of estimation 385
Standard normal distribution 346
tables of the, 346
Standard normal variable 346
Standard units 141
Standard values 141
Standardization 141, 345
Statistical control 125
Statistical generalization 382
Statistical hypothesis testing 407
Statistical inference 5, 280, 382, 583 (*see also* Inference)
Statistical regularity 284
Stem-and-leaf display 150 (*see also* Back-to-back stem-and-leaf display, Extended stem-and-leaf display)
Stochastic effect model 533, 534
Stochastic experiment 284, 291, 293, 302, 318
Stochastic variable 317, 351 (*see also* Continuous stochastic variable, Discrete stochastic variable)
cumulative distribution function of a, 324
domain of the, 319
mean of a, 329
notation for a, 318
standard deviation of a, 329
theoretical distribution of a, 329
Straight line 165
Stratum 25 (*see also* Subgroup)
Student's t distribution 365
Sturge's rule 58
Subgroup 25, 32 (*see also* Cluster, Stratum)
Subjective allocation 284
Success 332, 333, 340
Sum of squares 516, 529 (*see also* Block sum of squares, Error sum of squares, Treatment sum of squares)
treatment, 511, 521
mean treatment, 512, 522
Symmetry 97–98, 146–147 (*see also* Point of symmetry)
measures of, 123

t distribution 365–367
normal approximation of the, 367
tables of the, 366
Target population 18
Test for symmetry 589
Test for the homogeneity of independent samples 562
Test statistic 414, 422, 508 (*see also* Sign test statistic)
Testing *see* Lower-tailed testing, Two-tailed testing, Upper-tailed testing
Theoretical distribution 317, 319 (*see also* Binomial distribution, Cumulative distribution, Exponential distribution, Hypergeometric distribution, Normal distribution, Poisson distribution)
Theory of probability *see* Probability theory
Three-dimensional bar chart 76
Three-dimensional pie chart 81
Three-point method 200
Three-way contingency table 576
Ties 594, 597, 606, 611
Time intervals 337
Time points 219
Time reversal test 271, 272
Time scale values 225
Time series 219 (*see also* Seasonally adjusted time series)
Time series data 55
Total deviation 176
Total probability 297
Total sum of squares 511, 521, 529
Total variation 176
Transformation 225, 226 (*see also* Double-log transformation, Exponential curve, Hyperbolic curve, Inverse transformation, Single-log transformation)
Treatment effects 522
Treatments 509
Tree diagram, 294, 295, 303
Trend 220, 223, 250
Trial 332 (*see also* Bernoulli trial)
Trivariate data set 164, 202
Trivariate linear model 13, 202
Tukey 150, 155
Tukey's method of pairwise comparison 519
Two-sample case 419, 428
Two-sample problem 371, 594
Two-stage experiment 302
Two-tailed testing 408
Two-way column percentage table 76
Two-way frequency distributions
graphic representation of, 72–76

Type I error 411
Type II error 411
Typical seasonal index 230
  calculation of, 231–236

Unbiased 384
Unexplained deviation 176
Ungrouped data 99, 124
  central moment of, 143
  measures of location for, 99
  measures of spread for, 124
  raw moment of, 143
Uniform distribution 342
Uniform model 323, 324, 342
Unimodal distribution 147
Univariate data set 164
Universal set 286
Unweighted composite price index 252–254
Upper-tailed confidence interval 392
Upper-tailed testing 408

Validity studies 40
Value indices 252 (*see also* Quantity indices, Simple composite value index, Simple value indices)
Value weights 257
Van der Waerden test 622
Variable (*see also* Dependent variable, Discrete stochastic variable, Explanatory variable, Independent variable, Stochastic variable)
Variance 132, 143 (*see also* Population variance; *see also under* Regression line)
  differences in, 441
  equal, 371
  hypothesis testing for, 426, 443
Variance analysis *see* Analysis of variance
Variation
  sources of, 509
  coefficient of, 140
Venn diagram 287–297

Weight series 255, 263
Weighted arithmetic mean 102 (*see also under* Relative prices)
  properties of the, 101
Weighted composite price index 252, 254, 255 (*see also* Fixed weight composite price index)
Weighted composite price indices 259 (*see also* Drobisch index, Edgeworth price index, Irving Fisher index, Laspeyres index)
Weights *see under* Consumer price index of South Africa
Whiskers 155
Wilcoxon rank sign test *see* Rank sign test
Wilcoxon rank sum test 594
Wilcoxon test for symmetry 589 *see also* Signed rank test

Zonal samples 34

Laymen are inclined to regard statistics as an elevated science – a field reserved for intellectuals. *Modern statistics in practice* puts paid to this myth. It is a useful addition to the existing scientific and technical literature in this field while serving as an introduction for laymen by copiously illustrating the practical application of statistics in all disciplines and any undertaking.

The Afrikaans version of the book is now in its fifth edition and the need for an English version has emerged. *Modern statistics in practice* is designed for use by students and lecturers at tertiary academic institutions. However, applications of statistics in everyday situations are addressed and illustrated by practical examples.

Research in virtually all disciplines is based on statistical techniques. Researchers often have little or no background in this field. Although these techniques cannot be applied without a thorough understanding of the purpose, meaning and limitations inherent in the field, this book offers sound background information for researchers who need to acquaint themselves with the basics so that they can understand the fundamentals of processing data that they have collected.

Years of experience as lecturers and consultants in statistics have equipped the authors to produce a book that is academically sound but also practically orientated. *Modern statistics in practice* is a valuable addition to the library of any student, lecturer, researcher or layman.

ISBN 0627019897 02/01
Cdt: 229.00 Cash: 206.10
MODERN STATISTICS N PRACTICE
22UIK 11/12/95

JL van Schaik
ACADEMIC

ISBN 0 627 01989 7
9 780627 019890